Towards 3D facial morphometry: facial image analysis applications in anesthesiology and 3D spectral nonrigid registration

THÈSE N^{O} 7936 (2017)

PRÉSENTÉE LE 18 AOÛT 2017

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE TRAITEMENT DES SIGNAUX 5

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Gabriel Louis CUENDET

acceptée sur proposition du jury:

Dr A. Schmid, président du jury
Prof. J.-Ph. Thiran, directeur de thèse
Prof. V. Blanz, rapporteur
Prof. H. K. Ekenel, rapporteur
Dr F. Fleuret, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2017

Success consists of going from failure to failure without loss of enthusiasm.
— Winston Churchill

A mes parents,
A Jeanne...

Acknowledgements

This thesis is the result of a long process, which spread over several years. This journey, made of both good and difficult moments, profoundly changed me. During these years, I learned a lot, not only on the scientific side, or about what it is to do research, but also about myself. That has been the result of a large number of interactions with all the persons I worked and lived with during these years. I would like to take the opportunity to thank them here.

First of all, I would like to thank my supervisor, Prof. Jean-Philippe Thiran. Of course, none of this would even have been possible if you had not accepted me as a PhD student in your lab. Thank you, Jean-Philippe, for your trust and your friendly support in the difficult moments.

I would like to thank the members of my thesis jury: Dr. Alexandre Schmid, the president of the jury, Prof. Volker Blanz, Dr. François Fleuret, and Prof. Hazım Kemal Ekenel, for having accepted to be part of that jury, for their valuable feedback on the manuscript, and for their insightful comments and questions during the defense. I also would like to thank the anesthesiologists who came to us with this project, Prof. Patrick Schoettker and Dr. Christophe Perruchoud, for the fruitful collaboration.

I first came to LTS5 as a master student and started working under the supervision of Anıl. Anıl, you have been a supervisor, a mentor, and you are now a good friend. In a sense, you have paved the road for the rest of us, in the *Face group*, and I know that has not been easy. I am grateful for it. Even after you left the lab, you took the time to proofread this entire thesis and to provide your detailed feedback on my work, as well as much needed encouragement. Thank you, Anıl, for everything you have done. Murat, we started at the same time, we hiked in Switzerland together, we traveled to California and Sicily together (of course, you traveled much more...), we both went to IBM Research for an internship, and we shared an office, countless parties, and most of our thoughts about our PhD life. We even managed to both defend our thesis in a one-day interval! We might very well be PhD-brothers. In any case, we have fought together and we did it! I only have one thing to say: *Thanks bro'!*

I was very lucky to work in a subgroup of LTS5: the *Face group*. Most of the work presented in this thesis is, to a certain degree, team-work. Thank you, Christophe, for trying hard to make me code in a decent way and with decent tools, for importing and supplying me with the best *Saigne* has to offer, and for your help and ideas. Thank you, Marina, Damien, and Saeed for your enthusiasm to collaborate and everything we shared. Thank you to our supervisors, past and present, Hua and Hazım. It has been a pleasure being part and working in *the* Face group.

Of course, life in the lab is not limited to the Face group, and not even to the LTS5, and I would like to thank Sasan, Dimitri, Christophe, Vijay, Marina, and Murat for never saying no to a

table football break, my office mates Alia, Vijay, Marina, Murat and Anıl, my travel companions to Japan Franck, Devis and Emmanuel, my sailing buddy Tom, the diffusion group lead by Alessandro, Alia, Alessandra, Elda, Anna, Muhamed, David, Gab, and everyone that crossed path in the "corridor" Ricardo, Christina, Didrik, Mario, Sibylle, Carlos, Chris Paccolat, Adrien, Hamed, Eleni, and those who I forgot... A special mention to Sasan for the unhealthiest tennis game ever, and one of the best laugh. That was really fun, thanks Sasan! Maintaining (some) order in such an environment is not an easy job, so thank you Rosie for your efforts to keep everything under control and for always being so helpful with administrative problems.
During my six-months internship at IBM Research in Rüschlikon, I had the privilege to work with great people: thank you Maria for making me realize that, indeed, the process matters. Thank you, Peter, Marianna, Elina, Bogdan, Erwan and the system biology group for everything I learned during these six months. Of course, this period would not have been as fun without all the C225 occupants and pool opponents: Jari, Paul, Ondrej, Panos, Steffen, Marcel, Benedikt and last but not least, the *last survivors*: Alessandro, Mathieu and Simone.
If I spent all these years at EPFL without getting a yellow inventory sticker (which all furniture and equipment have), that is probably because I also had opportunities to escape and play music with great friends. Thank you, Andrea and Carlo, for all the adventures with our *trio Chromatique*. Thank you, Olivier and Thomas, for all the fun and the good work. Thank you, Stefan, for everything you taught me during this period, and not only musically, for being so true, and for the comforting discussions in tough times. You helped me find gold, literally, in the river, but not only.
Besides music, I am very lucky to have good old friends to share a few beers, go hike and ski, play petanque, and have fun with. Thank you, Mina, Ben, Daniel, Julien, Michael and Denis. Special thanks to Daniel for proof-reading the most important parts of this thesis, your help was much appreciated.
Finally I would like to thank my family-in-law and my family. There is nothing better than a whole Sunday playing cards with you to clear one's mind. Thank you, Antoine, Anaïs, Adrienne, Valentin, Matthieu, Juliana, Elisabeth and Roby. Elisabeth you have been incredibly supportive, especially at the end of the thesis, taking great care of the most important persons in my life when I was less available for them. You knew how to be there without imposing yourself, we owe you a lot, thank you. Anne, Matthieu, *Maman* and *Papa*, I feel very lucky to have such a family. Not only were you always extremely supportive, you also always showed interest in my work. *Maman* and *Papa*, I am forever grateful for everything you have done for us, and for the education you gave us. It opened quite some doors.
There is one person without whom I would never have been able to complete this long journey. Jeanne, you helped me in every possible aspect of this thesis; from supervising the data collection at the hospital to looking for solutions to help me manage the stress, you supported me, and stood me. Thank you for your incredible patience during these five years, for your love, and for everything you took on yourself, especially during this eventful past year.

Lausanne, 18th of August 2017 Gabriel Cuendet

Abstract

In anesthesiology, the detection and anticipation of difficult tracheal intubation is crucial for patient safety. When undergoing general anesthesia, a patient who is unexpectedly difficult to intubate risks potential life-threatening complications with poor clinical outcomes, ranging from severe harm to brain damage or death. Conversely, in cases of suspected difficulty, specific equipment and personnel will be called upon to increase safety and the chances of successful intubation.

Research in anesthesiology has associated a certain number of morphological features of the face and neck with higher risk of difficult intubation. Detecting and analyzing these and other potential features, thus allowing the prediction of difficulty of tracheal intubation in a robust, objective, and automatic way, may therefore improve the patients' safety.

In this thesis, we first present a method to automatically classify images of the mouth cavity according to the visibility of certain oropharyngeal structures. This method is then integrated into a novel and completely automatic method, based on frontal and profile images of the patient's face, to predict the difficulty of intubation. We also provide a new database of three dimensional (3D) facial scans and present the initial steps towards a complete 3D model of the face suitable for facial morphometry applications, which include difficult tracheal intubation prediction.

In order to develop and test our proposed method, we collected a large database of multimodal recordings of over 2700 patients undergoing general anesthesia. In the first part of this thesis, using two dimensional (2D) facial image analysis methods, we automatically extract morphological and appearance-based features from these images. These are used to train a classifier, which learns to discriminate between patients as being easy or difficult to intubate. We validate our approach on two different scenarios, one of them being close to a real-world clinical scenario, using 966 patients, and demonstrate that the proposed method achieves performance comparable to medical diagnosis-based predictions by experienced anesthesiologists.

In the second part of this thesis, we focus on the development of a new 3D statistical model of the face to overcome some of the limitations of 2D methods. We first present EPFL3DFace, a new database of 3D facial expression scans, containing 120 subjects, performing 35 different facial expressions. Then, we develop a nonrigid alignment method to register the scans and allow for statistical analysis. Our proposed method is based on spectral geometry processing and makes use of an implicit representation of the scans in order to be robust to noise or holes in the surfaces. It presents the significant advantage of reducing the number of free

parameters to optimize for in the alignment process by two orders of magnitude. We apply our proposed method on the data collected and discuss qualitative results.
At its current level of performance, our fully automatic method to predict difficult intubation already has the potential to reduce the cost, and increase the availability of such predictions, by not relying on qualified anesthesiologists with years of medical training. Further data collection, in order to increase the number of patients who are difficult to intubate, as well as extracting morphological features from a 3D representation of the face are key elements to further improve the performance.

Key words: 2D/3D facial image analysis; Difficult intubation prediction; 3D facial expressions database; 3D nonrigid registration; Computational geometry; Spectral mesh processing.

Résumé

En anesthésie, la détection et l'anticipation de l'intubation trachéale difficile sont cruciales pour la sécurité des patients. Lorsqu'il subit une anesthésie générale, un patient qui est, de façon non anticipée, difficile à intuber risque des complications pouvant mettre sa vie en danger et pouvant avoir des conséquences cliniques dommageables allant de douleurs sévères à des dommages cérébraux ou la mort. A l'inverse, dans les cas de difficulté anticipée, on fera appel à du personnel et un équipement spécifique afin d'augmenter la sécurité et les chances de succès de l'intubation.

La recherche scientifique en anesthésie a associé un certain nombre de caractéristiques morphologiques du visage et du cou avec un risque accru de difficulté d'intubation. Détecter et analyser ces caractéristiques, et potentiellement d'autres, de façon à prédire la difficulté d'intubation de manière robuste, objective et automatique peut donc améliorer la sécurité des patients.

Dans cette thèse, nous présentons d'abord une méthode pour classifier automatiquement des images de la cavité buccale en fonction de la visibilité de certaines structures oropharyngeales. Cette méthode est ensuite intégrée dans une nouvelle méthode complètement automatique, basée sur l'analyse d'images frontales et de profil du visage, pour prédire la difficulté d'intubation d'un patient. Nous fournissons également une nouvelle base de données de scans tridimensionnels (3D) du visage et présentons les étapes initiales menant à la création d'un modèle 3D complet du visage pouvant servir dans des applications de morphométrie faciale, parmi lesquelles la prédiction de la difficulté d'intubation trachéale.

Afin de développer et de tester notre méthode, nous avons récolté une importante base de données d'enregistrements multi-modaux de plus de 2700 patients ayant subi une anesthésie générale. Dans la première partie de cette thèse, nous extrayons automatiquement des caractéristiques morphologiques et d'apparence à l'aide de méthodes bidimensionnelles (2D) d'analyse d'image faciale. Ces caractéristiques sont utilisées pour entraîner un classificateur apprenant à discriminer les patients difficiles à intuber des patients faciles à intuber. Nous validons notre approche dans deux scénarios différents, dont un proche d'un scénario clinique réel, en utilisant 966 patients et démontrons que la méthode proposée atteint des performances comparables aux prédictions d'anesthésistes expérimentés, basées sur des diagnostiques médicaux.

Dans la seconde partie de cette thèse, nous nous concentrons sur le développement d'un nouveau modèle statistique en 3D du visage afin de surmonter les limites des méthodes 2D. Nous présentons d'abord EPFL3DFace, une nouvelle base de données de scans d'expressions

faciales en 3D contenant 120 sujets réalisant 35 expressions faciales différentes. Ensuite, nous développons une méthode d'alignement non-rigide afin de mettre les scans en correspondance et de permettre une analyse statistique. La méthode que nous proposons est basée sur le traitement de géométrie spectrale et utilise une représentation implicite des scans de façon à être robuste au bruit ou aux trous dans les surfaces 3D. Cette méthode présente l'avantage significatif de réduire de deux ordres de grandeur le nombre de paramètres à optimiser dans le processus d'alignement. Finalement, nous appliquons notre méthode sur la base de données récoltée et discutons les résultats qualitatifs.
Au niveau de performance actuel, notre méthode automatique de prédiction de l'intubation difficile a déjà le potentiel de réduire les coûts et d'augmenter la disponibilité de ces prédictions en ne dépendant pas d'anesthésistes qualifiés avec plusieurs années de formation médicale. Une plus ample collecte de donnée, de manière à augmenter le nombre de patients qui sont difficiles à intuber, ainsi que l'extraction de caractéristiques morphologiques à partir d'une représentation en 3D du visage sont les éléments clés afin d'améliorer encore les performances.

Mots clefs : Analyse d'image faciale 2D/3D ; Prédiction de l'intubation difficile ; Base de données 3D d'expressions faciales ; Alignement non rigide en 3D ; Géométrie computationnelle ; Traitement spectral de maillage 3D.

Contents

List of Figures

List of Tables

List of abbreviations

1D one dimensional

2D two dimensional

3D three dimensional

AAM active appearance model

ASM active shape model

AUC area under the curve

CED cumulative error distribution

CLM constrained local model

FN false negative

FNR false negative rate

FP false positive

FPR false positive rate

GPU graphics processing unit

HOG histogram of oriented gradients

ICP iterative closest point

IDS intubation difficulty scale

LBF local binary features

MHT Manifold Harmonics Transform

MPU multilevel partition of unity

MSE mean square error

NICP nonrigid iterative closest point

PCA principal component analysis

RBF radial basis function

RMSE root mean square error

ROC receiver operating characteristic

SDM supervised descent method

SIFT scale-invariant feature transform

SLAM simultaneous localization and mapping

SVD singular value decomposition

SVM support vector machine

TMD thyromental distance

TN true negative

TNR true negative rate

TP true positive

TPR true negative rate

TSDF truncated signed distance function

Introduction

Context and motivation

In the beginning of July 1966, in the artificial intelligence lab at the Massachusetts Institute of Technology (MIT), the "summer vision project" [Papert, 1966] was intended to mimic the human visual system by attaching a camera to a computer and having it recognize objects in a scene. The initial plan was to complete the project over the summer.

45 years later, as observed in [Szeliski, 2011], despite all the advances in the field of computer vision, the dream of having a computer interpret and understand an image at the same level as a two-year old, for example counting all the animals in a picture, remains elusive. Computer vision is difficult partially because it is an inverse problem. From a set of limited observations, such as pixels in an image, we want to infer the three dimensional (3D) nature of the imaged object. The available data, i.e. the pixels in the image of the object, contain insufficient information to fully specify the solution, as the 3D structure of the object was projected to a two dimensional (2D) representation. In order to constrain the solution, probabilistic models are used to disambiguate between potential solutions. Intuitively, that means that the 3D structure of the object can be recovered, if we have additional information about the object, for example if we know that it is a human face and have a model for human faces. During these 50 years, a number of methods and applications attracted the interest of the research community and led to certain successes. We refer the reader to the inspiring introductory chapter of [Szeliski, 2011] for a chronological review of computer vision's advances in the past decades, starting in the 1970s with pictorial structure [Fischler and Elschlager, 1973], edge detection [Davis, 1975], feature-based stereo correspondence algorithms [Dev, 1974, Marr and Poggio, 1976, Moravec, 1977, Marr and Poggio, 1979], or optical flow [Horn and Schunk, 1981, Huang, 1981, Lucas and Kanade, 1981], to the 2000s and the application of advanced machine learning techniques to large scale computer vision problems.

Since the early days, the human face has always been of great interest to computer vision researchers [Sakai et al., 1972]. Already in the 1970s, in [Fischler and Elschlager, 1973], the reported experiments have human faces as their subject, for three reasons: the availability of a set of pictures containing faces, the need for a single reference face that can be used on the complete dataset, and the fact that we are familiar with the human face, which facilitates the evaluation of performance. These reasons have remained valid and might explain that many

successes in computer vision are linked to works on the human face: the availability of facial images has never been so high on social media and the Internet in general, the human face exhibits relatively small variation, as compared to other object categories, and thus can be detected and modeled relatively accurately, and the human face has a crucial importance in social interactions and conveys a large amount of information about a person's state of mind and intentions.

As an example of great success, detecting faces in images, i.e. face detection, is considered solved in many settings, and real-world applications of face detection have spread widely since the seminal work of Viola and Jones [Viola and Jones, 2001] on boosting based face detection, which was the first algorithm that made face detection practically feasible in real-world applications. Today, the majority of the commercial digital cameras have an embedded face detector, allowing the camera to auto-focus. Since the beginning of the 2000s, applications in automatic facial image analysis flourished and include but are not limited to face recognition and verification [Zhao et al., 2003], face tracking for surveillance [Kalal et al., 2010], facial behavior analysis [Pantic and Rothkrantz, 2000], facial attribute recognition [Kumar et al., 2009], i.e. gender/age recognition [Fu et al., 2010] or assessment of beauty [Bottino and Laurentini, 2010, Zhang et al., 2017], face relighting and morphing [Yang Wang et al., 2009], facial shape reconstruction [Blanz and Vetter, 1999], as well as image and video retrieval. As a second example of computer vision success linked to the human face, face recognition has recently been reported to reach close to human-level performance [Taigman et al., 2014].

In very recent years, the usage of facial image analysis methods is on the rise in areas such as marketing and emotion analysis [Ahn and Picard, 2014, Ringeval et al., 2015], face-tracking systems to increase safety in cars [Dong et al., 2009, Gao et al., 2014], as well as in medicine [Baynam et al., 2011, Claes et al., 2012, Zhao et al., 2013, Kosilek et al., 2015], to name just a few. Facial landmarks detection and tracking is an extremely active field and recent progresses allow for fast and robust face trackers [Cao et al., 2012, Kazemi and Sullivan, 2014, Xiong and De la Torre, 2015, Ren et al., 2016]. These can detect and interpret specific features of the face, based on landmark positions, making them suitable for facial morphology analysis, or facial morphometry.

In this thesis, we focus on medical applications, in anesthesiology, of facial image analysis and, more specifically, facial morphometry. Prior to an operation which requires the patient to be under general anesthesia, the priority of the anesthesiologist, after having induced general anesthesia, is to ventilate the patient and secure his airways. Indeed, the patient is under the influence of drugs, whose main effects are the loss of consciousness, analgesia and muscular paralysis, and is unable to breath by himself. A standard way to enable mechanical ventilation is by introducing a tube in the trachea of the patient, through the vocal chords. This routine medical act is called tracheal intubation. For a large majority of patients, tracheal intubation does not present any difficulty, but for less than 10% of the patients, tracheal intubation can be difficult and put the patient at risk. Thus, detection and anticipation of difficult airway in the preoperative period is crucial for patients' safety. Research in anesthesiology have associated

a certain number of morphological features of the face and neck with higher risk of difficult intubation.

Detecting and analyzing these features, and potentially others, in order to predict the difficulty of tracheal intubation in a robust, objective, and automatic manner can therefore improve patients' safety. In the first part of this thesis, we thus describe advanced 2D facial image analysis methods to detect morphological features related to difficult intubation, hypothesizing that they could improve the prediction of difficult intubation. We demonstrate that the proposed method yields performance similar to state-of-the-art multifactorial tests performed manually by experienced anesthesiologists but does not require any measurement on the patient other than frontal and profile photographs, making it practical even for untrained personnel.

Nevertheless, the 2D methods used in this part suffer from limitations due to the loss of information happening during the 3D to 2D projection. In order to infer information that is contained in the 3D morphological structure of the face and neck, such as, hypothetically, the difficulty of intubation, first retrieving this 3D structure could help. Similarly to many inverse problems, this is only possible with strong priors about the structure, such as for example, a 3D face model. In the second part of the thesis, we focus on the first steps to build such a 3D face model, namely the recording of 3D facial surfaces of a population of 120 subjects, performing different facial expressions, and the nonrigid registration of these scans, such that statistical analysis can be applied.

In the next section, we detail the structure of this thesis and describe the relationship between its two parts and how chapters are interlinked with each other. Finally, the last section of this chapter lists the contributions of this thesis in a clear and succinct way.

Outline of the thesis

From a high level point-of-view, the core of this thesis is divided into two parts, each one presenting different research aspects of the same problem. In order to see clearly the link between these two parts and understand their relationship, one needs to keep in mind the focus of this work: predict the difficulty of intubation of patients using facial image analysis methods in order to improve the patients' safety.

Within that scope, part I presents different medical applications, in link with the prediction of difficult intubation: first, a Mallampati classification system, and second, a method for fully automatic prediction of difficult tracheal intubation. In these applications, key features, in terms of representation power for classification, are morphological features of the face. These morphological features are extracted from images using state-of-the-art 2D facial image analysis methods and, as such, suffer from limitations inherent to 2D methods. Specifically, these limitations are a high sensitivity to head pose variations and self occlusions and are due to the loss of information happening during the projection from the 3D world to the 2D image plane.

With respect to these limitations, a 3D method presents the advantage that, using a 3D model, the variation due to head pose is usually decoupled from the variation due to the identity and expression of the subject. Part II describes the initial steps required to build a 3D model of the face that could be used to extract morphological features in the scope of the prediction of difficult intubation. Specifically, these steps are the recording of a database of 3D scans of expressive faces from a variety of subjects and the re-parameterization of these scans into a common representation, which allows for statistical modeling. Ultimately, such a 3D statistical model of the face contains enough prior information about the 3D structure of face, such that this structure can be recovered even from new, unseen, and possibly self occluded, 2D images. The applications in which a 3D statistical model of the face is useful are not limited to the medical ones described in this thesis but also include expression recognition, mood or state-of-mind prediction, or visual speech recognition. As such, we are confident that the model resulting from this work will be useful in other applications as well.

At a finer level, this thesis is divided into chapters, which aim to be self-contained while presenting related aspects of this work. Chapter 1 is an introduction to 2D facial image analysis methods and provides background informations about the methods that were used throughout this thesis. Chapter 2 through chapter 4 constitute the part I described above while chapter 5 and 6 constitute part II. Finally, the last chapter, Conclusions, concludes this thesis and discusses some future perspectives. In the remaining of this section, we describe explicitly the contributions of each chapter and how they fit in the global scope of this thesis.

Chapter 1: Overview and benchmarking of 2D facial image analysis methods. This chapter describes a typical 2D facial image analysis pipeline and representative methods for face detection and facial landmark localization. It is intended as a smooth introduction to this field and a technical overview of the different categories of methods. Moreover, we perform a quantitative comparison of four well-known methods for facial landmarks localization on two different publicly available databases. In part I, these methods are essential tools used to extract morphological features from the face in the scope of the prediction of difficult intubation. In part II, they play a major role in the pre-processing of the 3D scans, namely in rigid registration.

Part I

Chapter 2: Introduction to the prediction of difficult tracheal intubation. This chapter specifically introduces the first part of this thesis. In this chapter, we first review some of the definitions of the difficult tracheal intubation and discuss their ambiguity. Then, we review existing automatic and manual methods of prediction of the difficult tracheal intubation and discuss their limitations. This chapter aims at providing a basic understanding of the difficult tracheal intubation prediction problematic to the reader without a medical background. A significant part of this chapter has been published, as introductory material, in [Cuendet et al., 2015] ©2015 IEEE.

Chapter 3: Automatic Mallampati classification. This chapter presents a method to classify images of patients, with the mouth wide open and the tongue protruding to its maximum, according to their modified Mallampati score, a simple indicator of potential difficulty to intubate. To the best of our knowledge, this is the first work proposing an automatic system to assess the modified Mallampati score from images. This work has been published in [Cuendet et al., 2012].

Chapter 4: Automatic prediction of difficult tracheal intubation. In this chapter of part I, we present a completely automatic method, based on facial morphometry, to predict a patient's difficulty of intubation with performance comparable to medical diagnosis-based predictions by experienced anesthesiologists. A large part of this chapter has been published in [Cuendet et al., 2015] ©2015 IEEE, and a patent is pending for this method [Schoettker et al., 2014].

Part II

Chapter 5: Background. This chapter introduces the second part of this thesis and provides some background about the different 3D methods used in this part. It first reviews existing databases of 3D facial scans and compares them to the new database of 3D facial expressions scans which we introduce in this thesis. It also provides a comprehensive description of the acquisition of 3D scans using a Microsoft Kinect®. Finally, spectral geometry processing methods on 3D meshes are introduced.

Chapter 6: Spectral nonrigid registration. The nonrigid registration of scans is the first step towards building a 3D statistical model from these. This chapter presents a novel 3D spectral nonrigid registration method and demonstrates its effectiveness on EPFL3DFace, a new database of facial expressions scans. A large part of this chapter has been submitted for publication in [Cuendet et al., 2017] and is currently under review.

Conclusions. To conclude this thesis, this chapter reviews key findings from our work and addresses future perspectives.

Contributions

The main contributions of this thesis are summarized below:

- A comprehensive overview of a selection of important methods in face detection and facial landmark localization with a comparative benchmark presenting quantitative results on two publicly available databases;

Contributions

- A Mallampati classification method, based on active appearance model (AAM) coefficients, trained and tested on images of the mouth cavity of 100 patients annotated by experienced anesthesiologists, which yields a high classification accuracy of 95% [Cuendet et al., 2012];
- A large database of facial images of patients, captured during the preoperative anesthesia consultation in two different hospitals
- A fully automatic method to predict patients' difficulty of intubation from facial images with performance comparable to those obtained by trained anesthesiologists [Schoettker et al., 2014, Cuendet et al., 2015];
- A large 3D facial expressions database, containing 35 different expressions performed by 120 subjects, suitable for a variety of applications in facial image analysis, such as expression recognition, mood detection, visual speech recognition, or 3D facial morphometry;
- A novel 3D spectral nonrigid alignment method using an implicit surface representation and a spectral embedding of the template as deformation model, thus reducing the number of free parameters in the optimization by a factor close to 100 [Cuendet et al., 2017].

1 Overview and benchmarking of 2D facial image analysis methods

The human face is the feature that allows us humans to easily recognize individuals. It conveys essential information about one's identity and, even amongst people we do not know, the face allows us to infer important characteristics such as their gender, their approximate age, or their origin or ethnicity. The human face is also an essential factor in physical attractiveness [Zhang et al., 2017].

The role of the human face in interpersonal communication is critical. If as much as two third of the communication between two people, or one speaker and a group of listeners, is indeed happening nonverbally, a large part of that nonverbal communication is conveyed by the face. The emotional state of a person and its intensity are communicated by the face, but also the behavioral intentions of that person. The relatively recent field of *Affective computing* is defined as computing that relates to, arises from, or influences emotions by Rosalind Picard in her seminal work [Picard, 1995]. In recent years, the advances in this interdisciplinary area at the frontiers of computer science, signal processing, wearable device technology, psychology, and neuroscience among many others have been enabled by using facial cues, among others. In verbal communication, and more specifically for speech recognition, the movements of the mouth are important cues in noisy conditions. By analyzing these movements, hearing-impaired people can even perform lip-reading. The eyes and the direction of the gaze provide information about where a person is porting his attention. As we will further discuss in this thesis, the human face can even provide information about one's health.

The identity of a person and the different characteristics linked to it, the expression of emotions, and the cues for nonverbal communication conveyed by the face can be captured visually. In that context, vision is a modality which is particularly cheap and easy to use: a simple camera records the information, it is non-invasive, and the visual information is continuously available (as opposed to audio for example, which could capture information only when the person is speaking).

Probably due to the importance of the aforementioned applications, as well as the extreme availability of face images, a lot of efforts have been put into developing better and faster

algorithms and methods to analyze face images. Facial image analysis has been an extremely popular research topic, in the last twenty years, and has been built on top of some of computer vision's greatest success stories.

In the scope of this thesis, these algorithms are essential tools for the medical applications presented and discussed in part I. Some of the methods introduced in part II, though handling three dimensional (3D) data, also make use of some of the algorithms introduced in this chapter. The information presented in this chapter also allows to better understand the advantages and limitations of each of the methods and will serve as background to motivate the development and the use of 3D models in part II of this thesis. Indeed, part II presents some work towards 3D models of the face in order to avoid limitations inherent to methods working on two dimensional (2D) images such as self occlusions or sensitivity to head pose variations.

This chapter thus provides a technical description of the most important elements of a typical 2D facial image analysis pipeline. We first describe such a pipeline from a high-level point of view in section 1.1. Section 1.2 then describes two popular methods for face detection. In section 1.3, we give some insights about the four main categories of face alignment methods for facial landmark localization. We describe representative approaches from each of these four categories, thus providing a comprehensive overview of existing face alignment methods. These methods are then benchmarked on publicly available datasets and the results of this benchmark are presented and discussed in section 1.4. Finally, we summarize and conclude this chapter in section 1.5.

1.1 Introduction

Let us consider that we have an image containing one or several human faces. Without loss of generality, we can consider only one face in the image and apply the same reasoning independently on each face when several faces are present. From that input, our goal is to automatically extract some information, which obviously depends on the application: if we are interested in facial recognition, we might want to extract the identity of the person present in the image, or in the case of age and gender classification, we would like to obtain the age and gender of that person. As described in part I of this thesis, we might also be interested in predicting whether performing tracheal intubation on that person might be difficult or not. For some applications, the temporal evolution of the information might also be important: for example, in facial expression recognition, we might be interested not only in the facial expression of the subject at one instant, but also in the temporal evolution of the subject's facial expression. The same reasoning applies for gaze tracking, where, as the name suggests, we would like to extract the temporal evolution of the direction of the gaze and not just one direction at one given moment.

At a high level, a typical pipeline in facial image analysis is thus composed of the following modules: face acquisition (face detection and facial landmark localization), feature extraction,

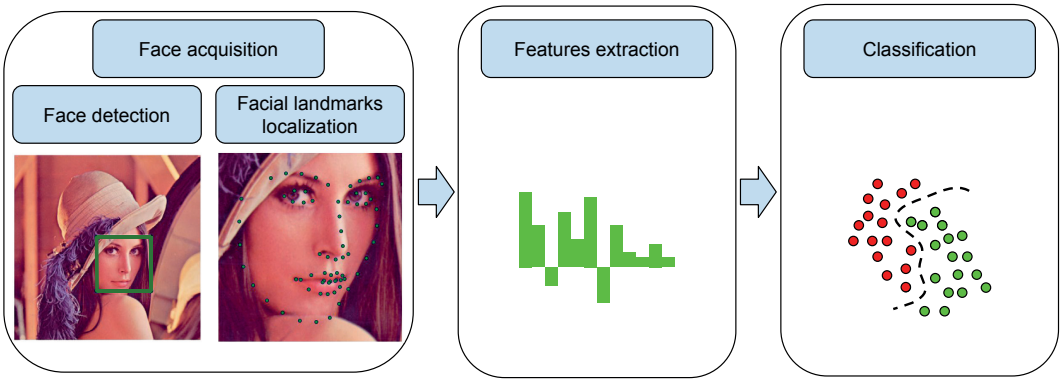


Figure 1.1 – Facial image analysis pipeline

and classification. Such a typical pipeline is illustrated in figure 1.1.

Face acquisition aims at the localization of the face in the image. This step is generally divided into two. First, a face detection step provides the rough location of the face in the image (if any). This provides the region of the image, generally as a rectangle, in which there is a face. Secondly, a facial landmark localization step aligns a known model of the face to the image with a face alignment method and allows to extract finer information, such as the locations of semantic landmarks on the face. This second step typically provides a set of facial landmarks, which are semantically meaningful, such as the corners of the eyes and mouth or the contour of the chin. The result is generally a vector containing the locations of the predefined landmarks in the image. These two tasks will be discussed in greater details in the next sections 1.2 and 1.3 of this chapter. Then, features are extracted. They can be of two types: either appearance based or geometric. Appearance based features are extracted from the texture of the image whereas geometric features are computed from the locations of the facial landmarks. This results in a new vector containing the features. Finally, relevant information is extracted from the feature vector using a classifier or a regressor if the desired output value is continuous. Upstream, this classifier needs to be trained from a large number of samples for which the ground-truth is available. It should also be noted that the dimension of the feature space can potentially be very high. In such cases, a dimensionality reduction algorithm can be beneficial, especially when the number of samples available for training the classifier is limited.

When the information that we are extracting from faces evolves over time and if we are interested in that temporal evolution, the input of the pipeline is generally not just one image anymore, but a sequence of images, such as a video. The face acquisition step is slightly different in that case and makes use of tracking to avoid having to detect the face if its position has not changed significantly. We will not go further into details about face tracking, as this is not relevant in the scope of this thesis. Instead, we focus on face detection and alignment in 2D images. Two methods for face detection are presented in section 1.2. The main classes of methods for facial landmark localization using face alignment are introduced in section 1.3.

1.2 Face detection

Given an image, it is necessary to first check for the presence of a face as well as its location and size in the image. A face detector thus takes the whole image as input and performs an exhaustive search in it. As the only available information is provided by the pixel values, i.e. the texture of the image, face detection is intrinsically an appearance-based method. Because of the lack of prior information about the location and scale of a face, the whole image needs to be processed. A very common strategy is to use a sliding window: a window whose size approximately corresponds to the size of the object to be detected, in our case a face, is slid at each position in the image successively across the whole image. For each position, the image is cropped to the size of that window and its content is considered as a candidate face. Either the raw pixels, or extracted features, are then fed to a classifier which outputs a decision about whether the content of the window is a face or not. As the size of the face is also unknown, the operation is generally repeated after down-sampling the image by different factors. This way, the size of the window remains the same, and thus the feature representation of a face does not change, but the image is processed at different scales. This is equivalent to looking for faces of different sizes. It should be noted that this strategy is common to many object detection methods and not just face detection. The problem of face detection is just a particular instance of the more generic problem of object detection and most of what will be introduced in this section applies to object detection in general. We choose to exemplify the process of object detection with faces as this is how these detectors are used in facial image analysis as well as in the scope of this thesis and because we hope that this will only make the description clearer to the reader.

As described above, the detection can be computationally expensive. A trade-off has thus to be found between simple features and classifiers, which might be fast but have difficulties to generalize and not be very accurate, and more accurate methods, which generalize better but are usually slower. Some of the challenges in face detection are the large changes in appearance introduced by head pose, facial expressions or occlusions. In order to be as robust as possible to these changes, a face detector generally requires several thousands of face and non-face example images for its training. Face detection is critical, as it is the first step of the pipeline. If no face is detected, no analysis can be performed.

In the remaining of this section we will first introduce the first real-time face detection method and probably the most used one, still today: the Viola-Jones face detector [Viola and Jones, 2001, Viola and Jones, 2004]. We then introduce a second very popular method, based on the idea of pictorial structures [Fischler and Elschlager, 1973]: the part-based detector of Yang and Ramanan [Yang and Ramanan, 2011, Yang and Ramanan, 2013]. These are the two face detectors that were used in the work presented in this thesis. Moreover, they lie at opposite ends of the spectrum of solutions in terms of trade-off between speed and accuracy: the Viola-Jones face detector is fast but not very accurate and not very robust against variations of head pose and facial expressions whereas the part-based detector of Yang and Ramanan is slow but can detect parts with a relatively high accuracy even with a lot of variations in

their relative positions. They are also representative instances of the two general schemes in face detection methods, as defined in [Zafeiriou et al., 2015]: Viola Jones is based on a rigid template, learned via a boosting based method, and the part-based detector of Yang and Ramanan is a deformable model that describes the face by its parts.

For a more complete review of the different methods that have been developed for face detection, we refer the reader to [Yang et al., 2002, Zhang and Zhang, 2010], and to the recent survey by Zafeiriou *et al.* [Zafeiriou et al., 2015].

1.2.1 Viola-Jones face detector

The Viola-Jones face detector, introduced in [Viola and Jones, 2001, Viola and Jones, 2004], describes a real-time method for face detection. An implementation of the Viola-Jones face detector is freely available[1] in the OpenCV library [Bradski, 2000].

The Viola-Jones face detector is defined by three main elements. The first one is the type of features that is used, which are reminiscent of Haar Basis functions. A new image representation called *integral image* allows to compute these features in a very efficient way, in *constant* time. The second key element of the method is the use of a variant of AdaBoost to perform feature selection and learn the classifier. The third key element is the use of a cascade of classifiers, which allows to speed up the classification of candidate regions. We will now discuss in more details each one of these three key components.

Haar-like features

There are good reasons to use features rather than the raw pixel intensities directly. The first one is that features can encode domain specific knowledge that would otherwise be difficult to learn from a limited quantity of training data. Features thus encode a higher level representation of the raw data. Of course, this also applies to other sorts of problems in machine learning, in a lot of different domains, and is not limited to face detection or even to image analysis. The second reason, more specific to that particular method, is that it is faster to process features rather than pixel intensities.

The Haar-like features used in the Viola-Jones face detector are basically differences between the sums of all pixels' values in different adjacent rectangular regions, as depicted in figure 1.2.

These features can be computed very efficiently using a novel representation of the input image called an *integral image*. A given pixel value in the integral image contains the sum of

[1] The open source OpenCV library can be downloaded from http://www.opencv.org

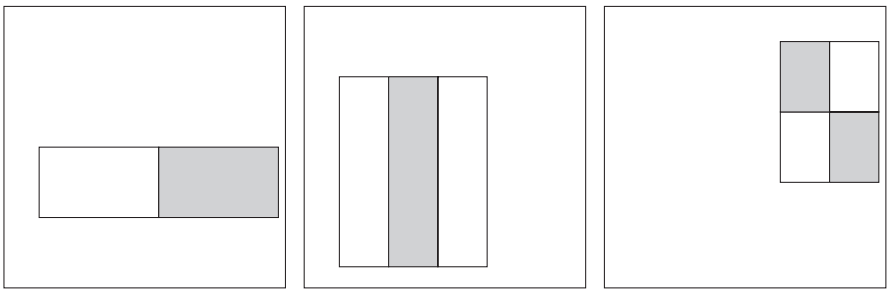


Figure 1.2 – Examples of Haar-like features shown relative to the detection window. For each example, the sum of the pixel values in the grey area is subtracted from the sum of pixel values in the white area. A different number of rectangles can be used: (a) 2 rectangles, either vertically (like this example), or horizontally stacked, (b) 3 rectangles, and (c) 4 rectangles. The size and position of each feature are different.

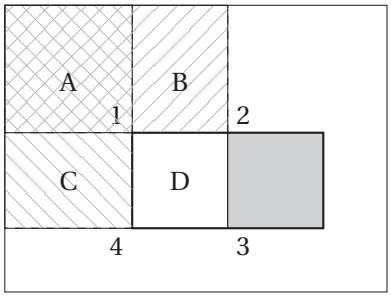


Figure 1.3 – The value of the integral image at location 1 is the sum of all pixel values in rectangle A. The value of the integral image at location 2 is the sum of all pixel values in rectangles A and B. Similarly the value of the integral image at location 4 is the sum of all pixel values in rectangles A and C. Finally, the value of the integral image at location 3 is the sum of all pixel values in rectangles A, B, C, and D. The sum of pixel values in the rectangle D can be computed as $\boldsymbol{I}_{\text{int}}(3) + \boldsymbol{I}_{\text{int}}(1) - \boldsymbol{I}_{\text{int}}(2) - \boldsymbol{I}_{\text{int}}(4)$, so accessing only four values of the integral image. The second half of the feature (in dark) is computed in a similar way and because the rectangles are adjacent the total number to access to the integral image is only six.

all the pixels above and to the left of that pixel in the original image (see eq. (1.1)).

$$\boldsymbol{I}_{\text{int}}(x,y) = \sum_{x' \leq x, y' \leq y} \boldsymbol{I}(x',y'), \tag{1.1}$$

where $\boldsymbol{I}_{\text{int}}$ denotes the integral image, $\boldsymbol{I}$ the original one, and (x, y) is the pixel position within the image.

Figure 1.3 illustrates how the sum of pixel values in any rectangle can be computed by reading four values in the integral image independently of the size of the rectangle. The complexity to compute the Haar-like features is thus constant.

Feature selection and classification with AdaBoost

The second key element of the Viola-Jones face detector is the feature selection and classification scheme using AdaBoost [Freund and Schapire, 1995]. Feature selection is critical as the number of features that are extracted from each detection window is much larger than the number of pixels in the corresponding image patch. For patches of size 24x24 pixels, thus containing 576 pixels, the exhaustive set of features is over 180000. The basis of the feature space is thus overcomplete and a small amount of these features can be combined to be discriminant. The role of feature selection is thus to find this limited set of discriminant features.

Once features have been extracted, both for positive examples (faces) and negative ones (non-faces), in principle any classifier could be used to learn a decision function. The Viola-Jones face detector uses the AdaBoost learning algorithm [Freund and Schapire, 1995]. The main idea is to combine a number of weak classifiers, whose accuracy does not need to be very good but just above random chance, in order to get a strong classifier. In addition, by enforcing that each weak classifier is using only one feature, keeping a limited number of weak classifiers T also performs feature selection, since these T classifiers are only using at most T features out of the complete set of features. A weak classifier $h_j(x)$ thus consists of one feature value ϕ_j, a threshold θ_j, and a parity value p_j indicating the direction of the inequality sign in equation (1.2).

$$h_j(x) = \begin{cases} 1 & \text{if } p_j\phi_j < p_j\theta_j \\ 0 & \text{otherwise} \end{cases}, \tag{1.2}$$

where x is a sub-window of an image from which the feature ϕ_j is extracted.

Cascade of classifiers

The third key element of the Viola-Jones face detector is the use of a cascade of classifiers in order to speed up the classification process. The idea is to first reject as many of the negative sub-windows as possible while retaining most of the positive instances, *i.e.* minimize false negatives. In the early stages of the cascade, simpler classifiers are trained using AdaBoost (see previous section). Their threshold is adjusted so as to minimize false negatives to reject the majority of sub-windows before more complex classifiers are used in subsequent stages to achieve low false positive rates. The numbers of features in the first five stages of the cascade are 1, 10, 25, 25, and 50. These first layers are thus very fast and allow to reject most of the negative sub-windows very early in the classification process. At each stage, only sub-windows classified as positive are passed to the next classifier. Negative results are directly rejected. This structure of the classification process can be seen as a degenerate decision tree [Fleuret and Geman, 2001].

In summary, the simplicity of the features and the weak classifiers, as well as the cascade

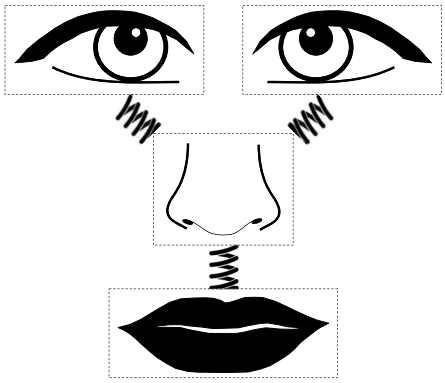


Figure 1.4 – Illustration of the concept of *pictorial structures*.

classification scheme of the Viola-Jones face detector, make it a relatively fast face detector. On the other hand, it is not very robust against large appearance changes due, for example, to occlusions, head pose variations, or large facial movements, such as opening the mouth wide or sticking the tongue out. In the next section, we describe another method for face detection which aims at modeling the face as an ensemble of parts, which can each have different appearances.

1.2.2 Parts based face detector

One fundamental limitation of the Viola-Jones face detector comes from its *holistic* representation of a face. As detailed in section 1.2.1, the face detector is trained to recognize a face as a whole, in any given sub-window. There are two main potential drawbacks with that holistic representation. The first one is that the global appearance of a face is impacted a lot if some parts of the face are occluded. The detector might thus be less robust to occlusions. The second drawback is that it is not possible to model separately different appearances corresponding to different head poses.

The pictorial structures representation introduced by Fischler and Elschlager [Fischler and Elschlager, 1973] provides a framework in which an object is modeled as a collection of parts arranged in a deformable configuration. Figure 1.4 illustrates the concept of pictorial structures. That framework is quite general, as it does not impose a particular appearance representation for the parts, neither does it specify the type of connections between parts. Felzenszwalb and Huttenlocher [Felzenszwalb and Huttenlocher, 2005] proposed a statistical formulation and efficient algorithms to learn pictorial structures from example images and match these to new unseen images, in order to use pictorial structures for object recognition.

A natural way to describe the arrangement of parts is by using an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertices of the graph $\mathcal{V} = \{v_1, ..., v_n\}$ are the parts and there is an edge $(v_i, v_j) \in \mathcal{E}$ for each pair of connected parts. A particular instance of the pictorial structure is then described by a configuration of parts $\mathcal{L} = \{l_1, ..., l_i\}$ where each l_i specifies the location, or the location and

orientation of the part v_i. Intuitively, given an image and a configuration of parts $\mathcal{L}$, the score s of that configuration in given by eq. (1.3).

$$s(\mathcal{L}) = \sum_{i \in \mathcal{V}} m_i(l_i) + \sum_{i,j \in \mathcal{E}} d_{ij}(l_i, l_j), \tag{1.3}$$

where $m_i(l_i)$ measures a local score as the degree of match when part v_i is placed at the location l_i in the image and $d_{ij}(l_i, l_j)$ is the score associated with the deviation of v_i and v_j from their expected locations and orientations.

In [Felzenszwalb et al., 2010], Felzenszwalb *et al.* present a complete object detector with discriminatively trained part-based models. The features used are histogram of oriented gradients (HOG) [Dalal and Triggs, 2010] and the classifier is a latent support vector machine (SVM).

With this formulation, the first limitation of an holistic representation has been addressed. Local parts are modeled independently and if one of them is occluded, but most of the others agree with the model, the score will still be high. This is, in a nutshell, the main advantage of local methods over holistic methods.

Yang and Ramanan further extended the idea of parts-based detector by adding a mixture of non-oriented pictorial structure [Yang and Ramanan, 2011] [Yang and Ramanan, 2013]. As mentioned at the beginning of this section, different head poses or facial expressions yield variations in appearance of the face. These variations are modeled separately in a mixture of pictorial structures associated with each part. The mixture component of part v_i is denoted t_i and called *type*. Types can thus model different expressions or different head poses. Moreover, it also becomes possible to model co-occurrence constraints that favor certain combinations of part types. Equation (1.4) describes the co-occurrence model.

$$s(\mathcal{T}) = \sum_{i \in \mathcal{V}} b_i^{t_i} + \sum_{i,j \in \mathcal{E}} b_{ij}^{t_i, t_j}, \tag{1.4}$$

where s is the score of the types $\mathcal{T} = \{t_i\}, i = 1, ..., n$, the first term $b_i^{t_i}$ favors particular types for each part, and the second term $b_{ij}^{t_i, t_j}$ favors particular co-occurrences of types t_i and t_j for parts v_i and v_j. As an example, on a lateral view of the face, all parts are viewed from the side and thus the types should all be those corresponding to a lateral view. It is not very likely that one part is viewed from the side and another one from the front.

Equation (1.3) thus becomes eq. (1.5).

$$s(\boldsymbol{I}, \mathcal{L}, \mathcal{T}) = \sum_{i \in \mathcal{V}} \boldsymbol{w_i^{t_i}} \cdot \boldsymbol{\phi}(\boldsymbol{I}, l_i) + \sum_{i,j \in \mathcal{E}} w_{ij}^{t_i, t_j} \cdot \psi(l_i - l_j) + s(\mathcal{T}), \tag{1.5}$$

where s is the score of a particular configuration $\mathcal{L}$ of types $\mathcal{T}$, on the image $\boldsymbol{I}$, $\boldsymbol{w_i^{t_i}}$ is a learned template for part v_i tuned for type t_i, $\boldsymbol{\phi}(\boldsymbol{I}, l_i)$ is a feature vector extracted from location l_i on the image, $w_{ij}^{t_i, t_j}$ is the spring template, with a given rest location and rigidity, for the pair of

parts v_i and v_j tailored for the types t_i and t_j, and $\psi(l_i, l_j)$ is the relative location of part v_i with respect to part v_j, modeled with a quadratic function.

With the mixture of parts, the second limitation of a holistic representation has also been addressed. Different head poses or facial expressions can now be modeled separately. In this thesis, we use an open-source implementation[2] of the parts-based detector described in [Yang and Ramanan, 2011].

In this section, we have introduced the first step of a facial image analysis pipeline, face detection, through two methods for face detection, the holistic Viola-Jones face detector and the local parts-based detector of Yang and Ramanan. We will not go into more details, even though there is a lot more to say, especially about training as well as about the practical implementation of the detectors. The main goal of this section is to provide insights about these two detectors in order to better understand the choices that were made later in the thesis.

1.3 Facial landmark localization

The second step of face acquisition, in the facial analysis pipeline (see fig. 1.1), is facial landmark localization. From the location of a face in the image provided by the face detector, as described in section 1.2, semantic facial feature points are detected. These facial landmarks correspond to fiducial facial parts, such as the corner of the eyes, the tip of the nose or the contour of the mouth. The number and the type of facial landmarks can vary and they mostly depend on the method and the application scenario. Generally speaking, localizing more facial landmarks provides richer information, although the detection becomes more time-consuming. Figure 1.5 shows a typical set of 68 facial landmarks. Those were defined in the Multi-PIE database [Gross et al., 2010] and later adopted for the annotations of the *300 Faces In-The-Wild Challenge* [Sagonas et al., 2015].

Facial landmark localization is an extremely active research area, with many related research topics and real-world applications. Despite the vast amount of methods that have been published in that area in the past 20 years, Wang *et al.* [Wang et al., 2014] proposed to group them into four categories, based on how the shape and appearance variations are modeled. These four categories are: active appearance model (AAM)-based methods, constrained local model (CLM)-based methods, regression-based methods, and other methods. Figures 1.6 and 1.7 show a timeline for the development of facial landmark localization methods in these four categories. As illustrated already with the two examples of face detector presented in section 1.2, there are two main approaches to model the appearance of a deformable object such as the face: local methods and holistic methods.

In this section, a number of methods will be introduced, spanning each one of these four

[2] This C++ implementation developed and maintained by Hilton Bristow, Willow Garage is available on Github: https://github.com/wg-perception/PartsBasedDetector

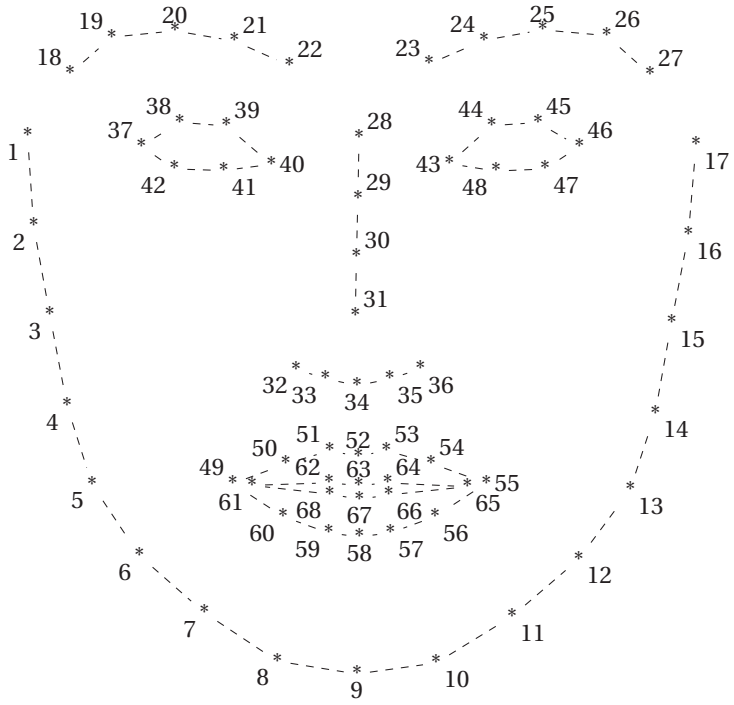


Figure 1.5 – 68 facial landmarks as defined in the Multi-PIE database [Gross et al., 2010]

categories. We do not aim at providing an exhaustive survey of face alignment methods but would like to provide a comprehensive overview of some of the most representative ones. For more complete recent surveys, we refer the reader to [Wang et al., 2014] and [Sagonas et al., 2015].

AAM-based methods use a holistic model of appearance. Moreover, both the shape variations and the appearance variations are represented as linear models and are usually coupled. These methods are described in subsection 1.3.1. CLM-based methods typically model the appearance variation locally around each facial landmarks independently using local experts. Each local expert allows to compute a response map around each facial landmarks. The facial landmark localizations are then predicted from these response maps, refined by a shape prior. CLM-based methods are described in subsection 1.3.2. More recently, regression-based methods have become very popular. They estimate the shape by learning a regression directly from the appearance to the facial landmarks. They do not define an explicit shape or appearance model. These methods, and more specifically the *supervised descent method* (SDM) of Xiong and De la Torre [Xiong and De la Torre, 2013], which is used throughout this thesis, are described in 1.3.3. Finally, the category of other methods contains graphical model-based, joint face alignment methods, independent facial feature point detectors, and deep learning-based methods.

The facial landmark localization problem can be described as an image alignment problem. In this thesis, we use these two terms interchangeably. Image alignment is the process consisting

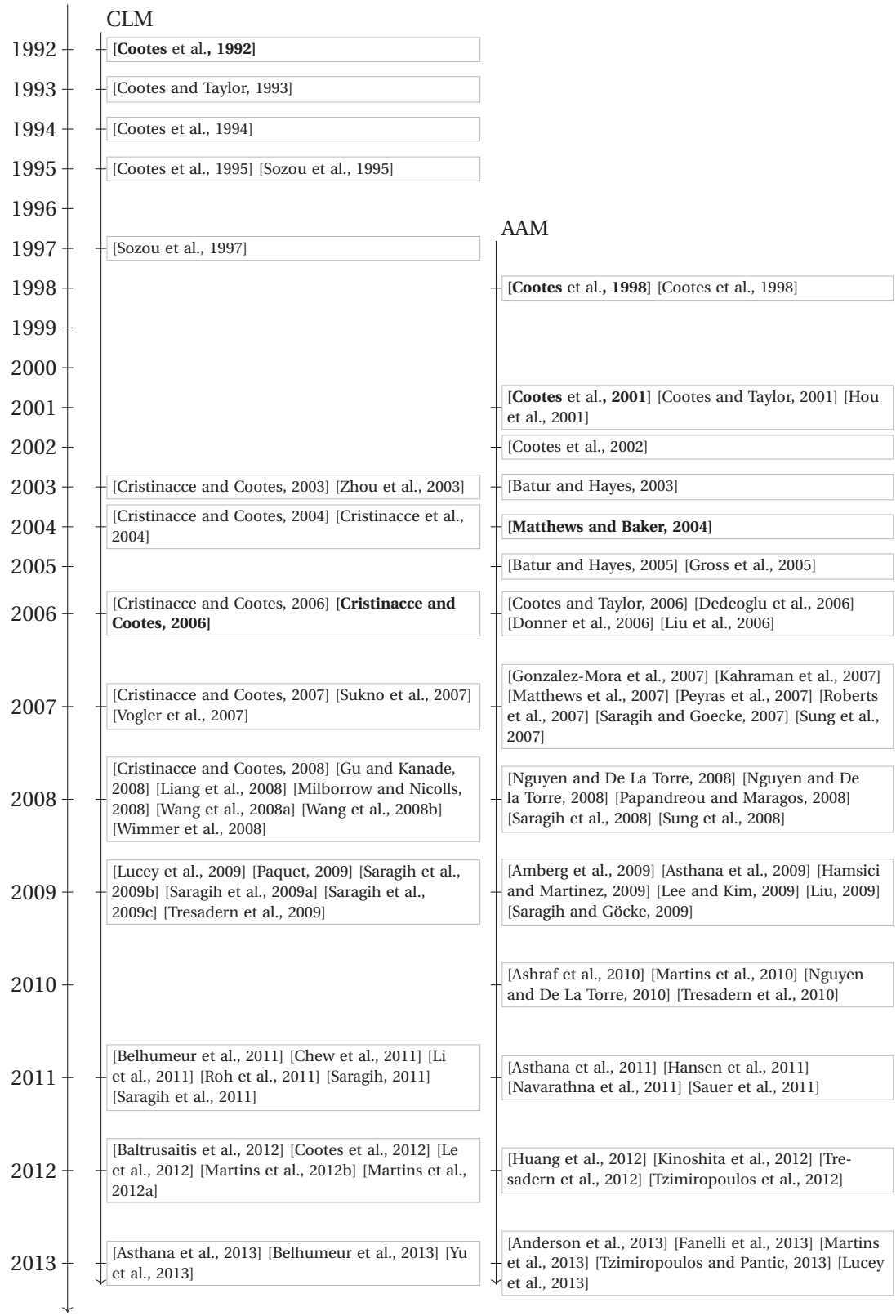


Figure 1.6 – Timeline of the development of CLM and AAM-based methods in facial landmark localization. This figure is based on an original figure from [Wang et al., 2014].

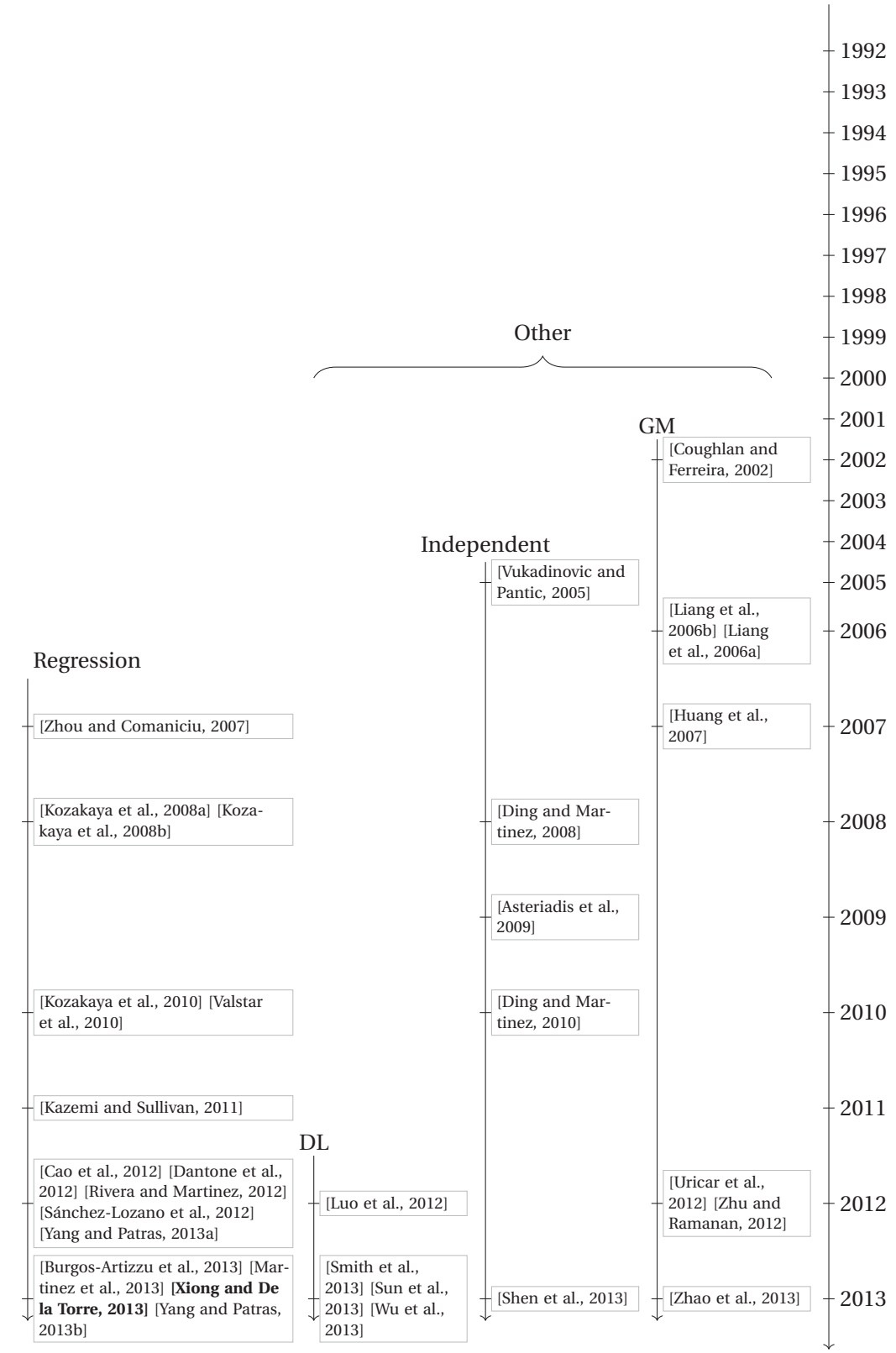


Figure 1.7 – Timeline of the development of Regression-based and other methods in facial landmark localization. This figure is based on an original figure from [Wang et al., 2014].

of rigidly moving and non-rigidly deforming a *template* to minimize its *distance* to a query image. Since Lucas and Kanade's seminal work [Lucas and Kanade, 1981], image alignment has become one of the most widely used techniques in computer vision. Its applications to faces include face fitting [Matthews and Baker, 2004], tracking [Black and Jepson, 1998] or face coding [Baker et al., 2004]. With the introduction of active shape models (ASMs) [Cootes et al., 1992] and active appearance models (AAMs) [Cootes et al., 2001] [Matthews and Baker, 2004] generative model-based face alignment has become very popular.

Image alignment process is characterized by three elements: *template representation*, *distance metric*, and *optimization scheme*. As examples, the template can be represented using a simple image patch, or the more sophisticated ASM or AAM, the mean square error (MSE) between the warped image and the template is one of the most widely used distance metrics and for the optimization, gradient descent methods are commonly used to iteratively update the shape parameters.

All these methods are supervised learning-based methods and thus can be decomposed into two phases: the learning phase and the fitting phase. During the learning phase, a large number of training images and ground truth facial landmark localizations are provided to the algorithm which learns its internal representations and parameters from those. During the fitting phase, the facial landmarks are localized on a previously unseen image on which only the location of the face is known, as provided by the face detector.

1.3.1 Active appearance models (AAM)

In the case of AAM, the template representation uses linear subspaces to model the object's shape and its shape-free appearance. *Combined* AAMs further model the correlation between the shape and appearance variations using an additional joint eigenspace, as described in the seminal work of Cootes *et al.* [Cootes et al., 1998]. *Independent* AAMs, on the other hand, consider two separate linear subspaces for the shape and the appearance. This presents advantages at fitting time and allows the use of the efficient *inverse compositional* method, as described in [Matthews and Baker, 2004].

Training Given a training set of face images, each image is manually annotated with a set of L 2D landmarks, $\{(x_i, y_i)\}, i = 1, ..., L$. These images are first rigidly aligned using Procrustes analysis [Goodall, 1991]. This step removes variations due to rigid transforms in the set of training shapes to keep only variations due to local, nonrigid shape deformation.

The collection of landmarks, or shape vector, $\boldsymbol{s} = (x_1, y_1, x_2, y_2, ..., x_L, y_L)^T$ of one image is treated as one observation from the random process defined by the shape model. AAMs model both the shape variations and appearance variations as linear models. The resulting model thus describes the shape, $\boldsymbol{s}$, as a linear combination of a reference shape and linear

bases as shown in equation (1.6).

$$\boldsymbol{s}(\boldsymbol{\alpha}) = \boldsymbol{s}_0 + \boldsymbol{P_s}\boldsymbol{\alpha} = \boldsymbol{s}_0 + \sum_{i=0}^{n} \alpha_i \boldsymbol{s}_i, \tag{1.6}$$

where $\boldsymbol{s}_0$ is a reference shape, $\boldsymbol{P_s} = \{\boldsymbol{s}_i\}$ is the matrix containing the set of orthonormal shape basis vectors $\boldsymbol{s_i}$, describing the modes of variation of the shape, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ are the shape parameters. The modes of variation are computed by performing principal component analysis (PCA) on the set of aligned training shapes and keeping the eigenvectors corresponding to the largest eigenvalues. The reference shape of the linear model, $\boldsymbol{s}_0$, is often the mean shape of the set of aligned training shapes.

In order to model the shape-free appearance, a warping function from the model coordinate system to the coordinates in the image observation is defined as $\boldsymbol{W}(x, y; \boldsymbol{\alpha})$, where (x, y) is a pixel coordinate within the face region defined by the mean shape $\boldsymbol{s}_0$ and $\boldsymbol{\alpha}$ are the shape parameters. This warping function is usually a piecewise-affine warp for each triangle in $\boldsymbol{s}_0$. We denote the resulting warped image as an N-dimensional vector $\boldsymbol{I}(\boldsymbol{W}(\boldsymbol{x}; \boldsymbol{\alpha}))$, where $\boldsymbol{x}$ is the set of all pixel coordinates within the mean shape $\boldsymbol{s}_0$. Given the shape model, each facial image is warped into the mean shape via the above warping function. Similarly to the shape model, PCA is applied to the set of shape-normalized appearances from all training images and the resulting model represents an appearance as described in equation (1.7).

$$\boldsymbol{A}(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{A}_0(\boldsymbol{x}) + \boldsymbol{P_a}\boldsymbol{\beta} = \boldsymbol{A}_0(\boldsymbol{x}) + \sum_{i=0}^{m} \beta_i \boldsymbol{A}_i(\boldsymbol{x}), \tag{1.7}$$

where $\boldsymbol{A}_0$ is the mean appearance, $\boldsymbol{P_a} = \{\boldsymbol{A}_i\}$ is the matrix containing the set of orthonormal appearance basis vectors $\boldsymbol{A_i}$, describing the modes of variation of the appearance, and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_m)^T$ are the appearance parameters.

The shape and appearance of a face can thus be described by the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Combined AAMs concatenate these two vectors into a single vector $\boldsymbol{b}$ as described by equation (1.8).

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{W_s}\boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W_s}\boldsymbol{P_s^T}(\boldsymbol{s} - \boldsymbol{s_0}) \\ \boldsymbol{P_a^T}(\boldsymbol{A} - \boldsymbol{A_0}) \end{pmatrix}, \tag{1.8}$$

where $\boldsymbol{W_s}$ is a diagonal matrix that weights each shape parameter to compensate for the difference in units between shape parameters and texture parameters. PCA is applied on b and results in eigenvectors $\boldsymbol{Q}$ and *combined appearance* parameters $\boldsymbol{c}$, controlling both the shape and texture parameters as shown in equation (1.9).

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{Q_s} \\ \boldsymbol{Q_a} \end{pmatrix} \boldsymbol{c}. \tag{1.9}$$

As the model is a linear model, the shape and the texture can be expressed directly as functions

of $\boldsymbol{c}$, as shown in equation (1.10).

$$\boldsymbol{s}(\boldsymbol{c}) = \boldsymbol{s}_0 + \boldsymbol{P_s W_s Q_s c}, \qquad \boldsymbol{A}(\boldsymbol{x};\boldsymbol{c}) = \boldsymbol{A}_0(\boldsymbol{x}) + \boldsymbol{P_a Q_a c}. \tag{1.10}$$

Fitting The fitting procedure is based on an *analysis by synthesis* approach. The intuition is to find the optimal parameters such that the synthesized image is as similar as possible to the observed image. Usually, the MSE between the warped observation and the synthesized appearance instance is used as the distance metric and the resulting cost function to minimize is described in equation (1.11).

$$J(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{N}\sum_{\boldsymbol{x}\in\mathscr{R}(\boldsymbol{s}_0)} \|\boldsymbol{I}(\boldsymbol{W}(\boldsymbol{x};\boldsymbol{\alpha})) - \boldsymbol{A}(\boldsymbol{x};\boldsymbol{\beta})\|^2 = \frac{1}{N}\sum_{\boldsymbol{x}\in\mathscr{R}(\boldsymbol{s}_0)} \|\boldsymbol{E}(\boldsymbol{x})\|^2, \tag{1.11}$$

where N is the total number of pixels within the face region $\mathscr{R}(\boldsymbol{s}_0)$ defined by the mean shape. The difference between the warped observation and the synthesized appearance instance is the error image, $\boldsymbol{E}(\boldsymbol{x})$. It should be noted that this cost function is generally noisy with a lot of local minima and no guarantee for the global minimum to be at the right location.

There are various approaches to minimize the cost function of equation (1.11). A natural way is to use a standard gradient descent optimization algorithm. This approach is very slow because the partial derivatives and gradient direction need to be recomputed at each iteration.

An alternative fitting approach is to linearize the relationship between the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the error described in equation (1.11). In practice, additive increments $\Delta\boldsymbol{\alpha}$ and $\Delta\boldsymbol{\beta}$ are computed as linear functions of the error image, $\boldsymbol{E}(\boldsymbol{x})$, as described in equation (1.12).

$$\Delta\alpha_i = \sum_{\boldsymbol{x}\in\mathscr{R}(\boldsymbol{s}_0)} R_{\alpha,i}(\boldsymbol{x})\boldsymbol{E}(\boldsymbol{x}), \qquad \Delta\beta_i = \sum_{\boldsymbol{x}\in\mathscr{R}(\boldsymbol{s}_0)} R_{\beta,i}(\boldsymbol{x})\boldsymbol{E}(\boldsymbol{x}). \tag{1.12}$$

The parameters are then updated in the following manner: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}$ and $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \Delta\boldsymbol{\beta}$. These additive increments are considered to be *constant* with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The update functions can be estimated by systematically perturbing the model parameters $\Delta\boldsymbol{\alpha}$ and $\Delta\boldsymbol{\beta}$ and saving the corresponding error image $\boldsymbol{E}(\boldsymbol{x})$. The values of $\boldsymbol{R_\alpha}(\boldsymbol{x})$ and $\boldsymbol{R_\beta}(\boldsymbol{x})$ are then estimated by linear regression.

Baker and Matthews [Baker and Matthews, 2004] showed that AAM-based image alignment algorithms can be classified by two criteria: how the updates are made, these are either *additive* or *compositional*, and in which reference frame the optimization is performed. The reference frame is either the model reference frame, in which case the image is warped to the reference shape $\boldsymbol{s}_0$ of the model, in a *forward* manner, or the image itself, in which case the generated texture is warped to the image in an *inverse* manner.

In order to be generic, we assume a warping function $W(\boldsymbol{x},\boldsymbol{p})$, parametrized by $\boldsymbol{p}$. Following an iterative approach as introduced above, at each iteration of the algorithm, the parameters are updated by $\Delta\boldsymbol{p}$ in order to improve the match. The optimization over the update $\Delta\boldsymbol{p}$ thus

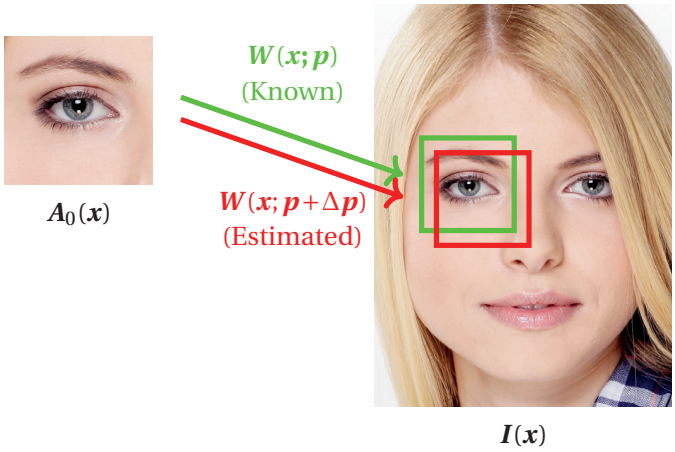


Figure 1.8 – Forward additive. This figure is based on an original figure from [Matthews and Baker, 2004].

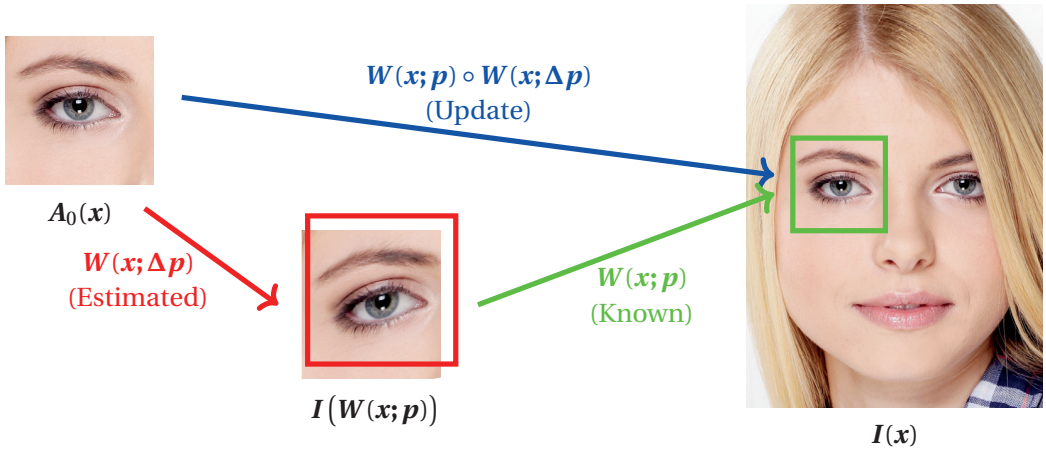


Figure 1.9 – Forward compositional. This figure is based on an original figure from [Matthews and Baker, 2004].

falls into one of the four following cases.

Forward additive Choose $\Delta \boldsymbol{p}$ to minimize equation (1.13) and update the parameters in the following way $\boldsymbol{p} \leftarrow \boldsymbol{p} + \Delta \boldsymbol{p}$. This is illustrated in figure 1.8.

$$J(\boldsymbol{p}) = \sum_{\boldsymbol{x} \in R(\boldsymbol{s}_0)} \| \boldsymbol{I}(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p} + \Delta \boldsymbol{p})) - \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}) \|^2. \tag{1.13}$$

Forward compositional Choose $\Delta \boldsymbol{p}$ to minimize equation (1.14) and update the parameters in the following way $\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p}) \leftarrow \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p}) \circ \boldsymbol{W}(\boldsymbol{x}, \Delta \boldsymbol{p})$. This is illustrated in figure 1.9.

$$J(\boldsymbol{p}) = \sum_{\boldsymbol{x} \in R(\boldsymbol{s}_0)} \| \boldsymbol{I}(\boldsymbol{W}(\boldsymbol{W}(\boldsymbol{x}, \Delta \boldsymbol{p}), \boldsymbol{p})) - \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}) \|^2. \tag{1.14}$$

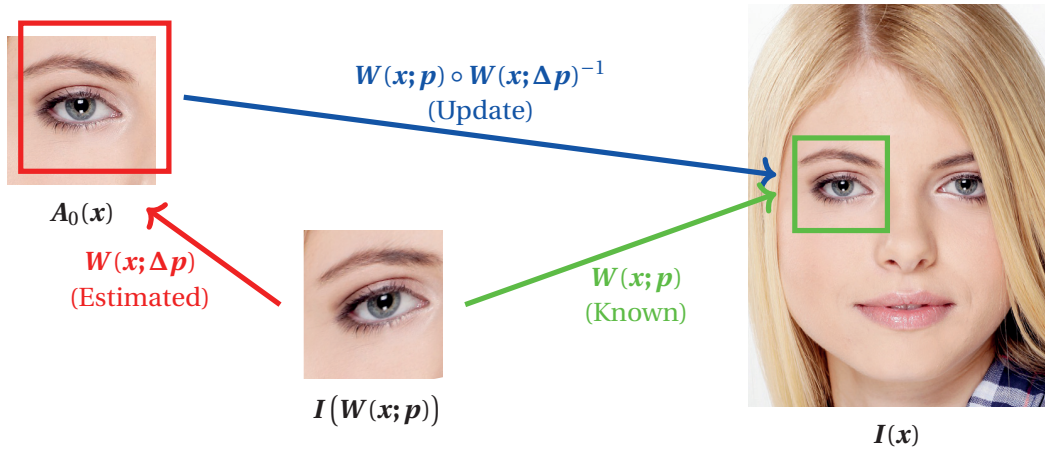


Figure 1.10 – Inverse compositional. This figure is based on an original figure from [Matthews and Baker, 2004].

Inverse additive Choose Δp to minimize equation (1.15) and update the parameters in the following way $p \leftarrow p + \Delta p$.

$$J(p) = \sum_{y \in R(s)} \| I(y) - A(W^{-1}(y, p + \Delta p), p) \|^2. \tag{1.15}$$

Inverse compositional Choose Δp to minimize equation (1.16) and update the parameters in the following way $W(x, p) \leftarrow W(x, p) \circ W(x, \Delta p)^{-1}$. This is illustrated in figure 1.10.

$$J(p) = \sum_{x \in R(s_0)} \| I(W(x, p)) - A(W(x, \Delta p), p) \|^2. \tag{1.16}$$

The Inverse Compositional AAM (AAM-IC) method proposed by Matthews and Baker [Matthews and Baker, 2004] greatly improves the performances by switching the role of the template and the input image which allows to precompute some of the parameters.

It has been observed that alignment performance of the AAM degrades quickly when they are trained on a large data set and fit to images that were not seen during the AAM training [Gross et al., 2005]. For a more complete overview of AAMs, we refer the interested reader to the very complete technical report by Cootes and Taylor [Cootes and Taylor, 2004].

1.3.2 Constrained Local Model (CLM)

CLM has been proposed by Cristinacce and Cootes [Cristinacce and Cootes, 2006]. The main difference with regards to AAM is the shape-free model of texture, which takes into account small patches around landmarks instead of the whole region defined by the mean shape. The template representation is thus not an *holistic* representation of the face anymore, but has a

more local character.

Training The shape model is exactly the same as for AAMs. The collection of annotated landmarks $\boldsymbol{s} = (x_1, y_1, x_2, y_2, ..., x_L, y_L)^T$ of one image is treated as one observation from the random process defined by the shape model described in equation (1.6). This shape model is learned by performing PCA on the aligned training examples, in the same way as for AAMs.

Unlike AAMs, the appearance of the local region around each landmark is modeled independently. These local patch experts are learned such that, when cross-correlated with an image region containing the corresponding facial landmark, they yield a strong response at the landmark location and weak responses everywhere else. This can be done either generatively [Cristinacce and Cootes, 2006] or discriminatively, by learning a classifier or a regressor [Cristinacce and Cootes, 2007].

Fitting CLM fitting, from a high level point of view, is defined as the search for the shape model parameters which jointly minimize the misalignment error over all landmarks, given by local patch experts, while regularized by the shape model as described in equation (1.17).

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} D_i(\boldsymbol{x}_i, \boldsymbol{I}) + R(\boldsymbol{\alpha}), \tag{1.17}$$

where D_i, the data term, is the misalignment error of landmark i and R is the regularization term, penalizing complex deformation of the shape.

The main challenges in optimizing efficiently equation (1.17) come from the data term D_i, which can have many local minima, as well as from potential outlying detections. A number of different strategies have been proposed to handle these challenges and optimize equation (1.17) efficiently.

If the location of the maximum in each response map $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ can be determined, the data term becomes simply the distance between each landmark and the location of the maximum in the corresponding response map and the shape is regularized with the norm of the shape coefficients vector in order to avoid outliers, as shown in equation (1.18).

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \boldsymbol{\mu}_i\|^2 + \|\boldsymbol{\alpha}\|^2, \tag{1.18}$$

where the weights w_i show the confidence over the location of the maximum in the i$^{\text{th}}$ response map.

For more details and for a review of different optimization strategies, we refer the reader to [Saragih et al., 2011].

1.3.3 Regression-based face alignment

Regression-based methods for face alignment were introduced relatively recently. One of the most influential early regression-based face alignment method is the Explicit Shape Regression by Cao *et al.* [Cao et al., 2012].Unlike AAM-based methods and CLM-based methods, in which a parametric shape model was trained, regression-based methods do not use any shape model. Instead, the locations of landmarks are inferred by directly learning a vectorial regression function from the image. The regressor is trained to minimize the alignment error over training data in a holistic manner in two respects. First, all facial landmarks are regressed jointly and second, the image features for each facial landmark are not necessarily computed from the local neighborhood around these landmarks. Moreover, boosted regression is often used in order to gradually converge towards the shape. The first regressors handle large shape variations but do not fit accurately and the last ones handle only small shape variations but provide higher accuracy. Thus, each regressor refines the localization of the landmarks by producing a vectorial update $\boldsymbol{\Delta s}$, which is added to the current landmark locations estimate. The high level idea of cascaded regression-based methods is summarized in algorithm 1.

Algorithm 1 Cascaded shape regression

Require: Image $\boldsymbol{I}$, initial shape $\boldsymbol{s}_0$
Ensure: Estimated shape $\boldsymbol{s}_T$
1: **for** $t = 1$ to T **do**
2: $\quad \boldsymbol{\phi}_t = \boldsymbol{\Phi}_t(\boldsymbol{I}, \boldsymbol{s}_{t-1})$ ▷ Compute shape-indexed features
3: $\quad \boldsymbol{\Delta s} = \boldsymbol{R}_t(\boldsymbol{\phi}_t)$ ▷ Apply regressor
4: $\quad \boldsymbol{s}_t = \boldsymbol{s}_{t-1} + \boldsymbol{\Delta s}$ ▷ Update shape
5: **end for**

The key differences between particular regression-based approaches are the type of features and the regressor. The type of features that are used as input to the regression and the regressor are interdependent.

In this chapter, we will introduce two recent regression-based facial landmarks localization methods: the supervised descent method (SDM) [Xiong and De la Torre, 2013] and the local binary features (LBF) method [Ren et al., 2014]. We use our own implementation of the SDM in both parts of this thesis, with some additional improvements over the original method, as described in [Qu et al., 2015]. Specifically, these improvements are a more robust regression, through the use of iteratively reweighted least squares (IRLS), RootSIFT features, and a coarse-to-fine fitting strategy and in-plane pose normalization during shape update. We also use our own implementation of the LBF, following the original paper [Ren et al., 2014].

Supervised descent method (SDM)

In order to minimize nonlinear least squares functions, Xiong and De la Torre introduce a supervised descent method (SDM) [Xiong and De la Torre, 2013] and show how it achieves state-of-the-art performances on facial landmarks localization.

Training The problem of localizing L facial landmarks $\boldsymbol{s} = (x_1, y_1, ..., x_L, y_L)^T$ in an image $\boldsymbol{I} \in \mathbb{R}^{m \times 1}$, can be posed as

$$f(\boldsymbol{s}_0 + \Delta\boldsymbol{s}) = \|\boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s}_0 + \Delta\boldsymbol{s}) - \boldsymbol{\phi}_*\|_2^2, \tag{1.19}$$

where $\boldsymbol{\Phi}$ is a feature extraction function and $\boldsymbol{\phi}_* = \boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s}_*)$ represents the features extracted from the locations of the manually labeled landmarks $\boldsymbol{s}_*$. Note that $\boldsymbol{\Phi}$ depends both on the image $\boldsymbol{I}$ and the facial landmark locations $\boldsymbol{s}$. Such features are referred to as *shape-indexed* features. The goal of the SDM is to learn descent directions that produce a serie of updates $\Delta\boldsymbol{s}_{t+1} = \boldsymbol{s}_t + \Delta\boldsymbol{s}_t$, starting from $\boldsymbol{s}_0$ and converging to $\boldsymbol{s}_*$, in the training data.

Specifically, if we assume that $\boldsymbol{\Phi}$ is twice differentiable, we can apply a second order Taylor expansion to equation (1.19) as shown in (1.20).

$$f(\boldsymbol{s}_0 + \Delta\boldsymbol{s}) \approx f(\boldsymbol{s}_0) + \boldsymbol{J}_f(\boldsymbol{s}_0)^T \Delta\boldsymbol{s} + \frac{1}{2}\Delta\boldsymbol{s}^T \boldsymbol{H}(\boldsymbol{s}_0)\Delta\boldsymbol{s}, \tag{1.20}$$

where $\boldsymbol{J}_f(\boldsymbol{s}_0)$ and $\boldsymbol{H}(\boldsymbol{s}_0)$ are the Jacobian and the Hessian of f evaluated at $\boldsymbol{s}_0$. In the following, we will omit $\boldsymbol{s}_0$ to improve the readability. To find the optimal update $\Delta\boldsymbol{s}$, equation (1.20) can be differentiated with respect to $\Delta\boldsymbol{s}$ and set to zero as shown in equation (1.21). The condition that $\boldsymbol{\Phi}$ be differentiable is necessary for the derivation but will be dropped.

$$\begin{aligned} \frac{\partial f(\boldsymbol{s}_0 + \Delta\boldsymbol{s})}{\partial \Delta\boldsymbol{s}} = \boldsymbol{J}_f + \boldsymbol{H}\Delta\boldsymbol{s} &= 0 \\ \iff \Delta\boldsymbol{s} &= -\boldsymbol{H}^{-1}\boldsymbol{J}_f = -2\boldsymbol{H}^{-1}\boldsymbol{J}_h^T(\boldsymbol{\phi}_0 - \boldsymbol{\phi}_*), \end{aligned} \tag{1.21}$$

where the chain rule was used to show that $\boldsymbol{J}_f = 2\boldsymbol{J}_h^T(\boldsymbol{\phi}_0 - \boldsymbol{\phi}_*)$, where $\boldsymbol{\phi}_0 = \boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s}_0)$. The first update can thus be seen as a projection of $\Delta\boldsymbol{\phi}_0 = \boldsymbol{\phi}_0 - \boldsymbol{\phi}_*$ onto the row vectors of matrix $\boldsymbol{R}_0 = -2\boldsymbol{H}^{-1}\boldsymbol{J}_h^T$. $\boldsymbol{R}_0$ is a descent direction and produces an update starting from $\boldsymbol{s}_0$ and converging to the annotations $\boldsymbol{s}_*$ in the training data.

The computation of this descent direction is impractical as it requires $\boldsymbol{\Phi}$ to be two times differentiable or to compute expensive numerical approximations for the Jacobian and the Hessian. Moreover, the optimal update is given as a function of the annotations $\boldsymbol{\phi}_*$, which are only known at training time, but not during fitting. In order to be able to use the descent direction during fitting, equation (1.21) is rewritten as a generic linear combination of the feature vector $\boldsymbol{\phi}_0$ and a bias term $\boldsymbol{b}_0$ as described in equation (1.22).

$$\Delta\boldsymbol{s} = \boldsymbol{R}_0\boldsymbol{\phi}_0 + \boldsymbol{b}_0. \tag{1.22}$$

Both $\boldsymbol{R}_0$ and $\boldsymbol{b}_0$ are learned during training.

Given a set of training images $\{I^i\}$ and corresponding landmark locations ground truth $\{\boldsymbol{s}_*^i\}$, $\boldsymbol{R}_0$ and $\boldsymbol{b}_0$ are learned by minimizing a linear least squares problem, which can be solved in closed form. For each image, the expected loss between the predicted and the optimal

landmark displacement is minimized under different initializations $\boldsymbol{s}_0^i$.

$$\arg\min_{\boldsymbol{R}_0,\boldsymbol{b}_0} \sum_{I^i} \sum_{\boldsymbol{s}_0^i} \|\boldsymbol{R}_0\boldsymbol{\phi}_0^i + \boldsymbol{b}_0 - \boldsymbol{\Delta s}_*^i\|^2. \tag{1.23}$$

The choice of the feature extractor $\boldsymbol{\Phi}$ is completely free. The operator does not need to be differentiable and can be non-linear. In the original formulation of the SDM, the features are scale-invariant feature transform (SIFT) features [Lowe, 2004] and Qu *et al.* propose a comprehensive comparison of features to use in the SDM [Qu et al., 2015]. They also propose to solve the linear least squares problem with a regularization term and to use Iteratively Reweighted Least Squares [Green, 1984] in order to be less sensitive to noisy data samples.

Fitting As it is unlikely that the SDM can converge in a single iteration, unless f is a quadratic function, the algorithm generates a sequence of updates, as described in algorithm 1. For each iteration, the fitting is extremely simple. It is a linear regression from the feature vector, computed at the landmark previous locations $\boldsymbol{s}_{t-1}$, to an update $\boldsymbol{\Delta s}_t$, which is added to the current landmark locations estimate, as described in equation (1.24).

$$\boldsymbol{s}_t = \boldsymbol{s}_{t-1} + \boldsymbol{R}_{t-1}\boldsymbol{\phi}_{t-1} + \boldsymbol{b}_{t-1}, \tag{1.24}$$

where $\boldsymbol{\phi}_{t-1} = \boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s}_{t-1})$ is the feature vector extracted at previous landmark locations.

It can be seen from equation (1.24) that the limiting factor, in terms of speed, is the computation of the features. The alignment itself is limited to a matrix multiplication. With that respect, simpler features, which are faster to compute, have the potential to speed up the fitting of regression-based methods. This is a key point of the next method presented, the LBF.

Local Binary Features (LBF)

Ren *et al.* propose to use a local approach, as opposed to the SDM holistic approach, based on local binary features [Ren et al., 2014] [Ren et al., 2016]. Their claim is that computing the features locally helps to reduce the size of the feature pool and avoid two issues caused by a large feature pool: the training costs to learn the most discriminative feature combination are too high and, more importantly, many features are noisy and hinder the learning process.

Moreover, the SDM's hand-crafted SIFT features are replaced by a set of local binary features, which are learned from training data. Learning the features from data is interesting in general, as it learns task-specific features. Once they have been learned, these are extremely fast to compute, as they are based on pixel differences. The authors report localization rates above 3000 fps, at fitting time, for the Multi-PIE 68 landmarks shown in figure 1.5.

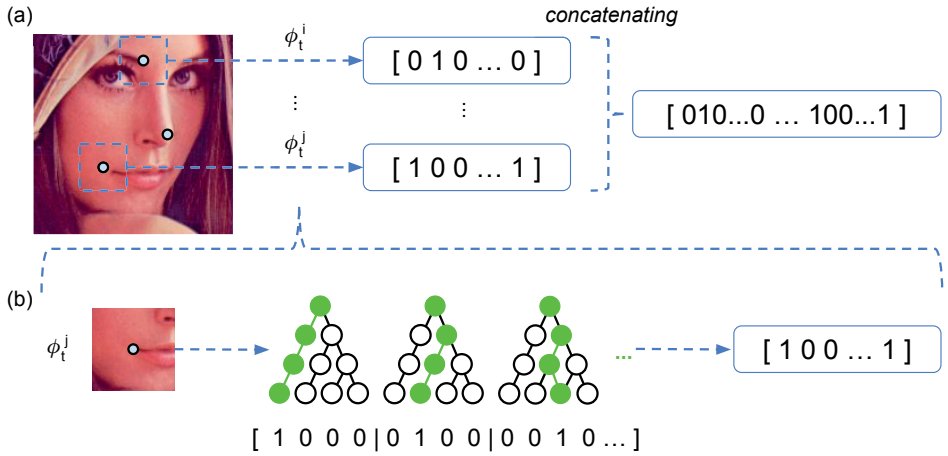


Figure 1.11 – Local binary features (a) Local feature mapping functions Φ_t^l encode the local region around each landmark into a binary feature vector; all local binary feature vectors are concatenated to form high-dimensional binary features. (b) Random forest are used as local mapping functions. This figure is inspired from Figure 2 in [Ren et al., 2014].

Training Both the linear regression matrix $\boldsymbol{R}_t$ and the feature mapping function $\boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s})$ are learned in two consecutive steps.

First, $\boldsymbol{\Phi}_t$ is decomposed into a set of per-landmark independent local feature mapping functions $\{\Phi_t^1, ..., \Phi_t^L\}$. Each one is learned by independently regressing the i^{th} landmark in the corresponding *local* region. All local features are concatenated into $\boldsymbol{\Phi}_t$. Then, $\boldsymbol{R}_t$ is learned by linear regression, similarly to the SDM. This two-stage process is repeated stage-by-stage in a cascaded fashion.

Each local feature mapping function is learned using a standard regression forest [Breiman, 2001], whose target is the ground truth shape increment, $\boldsymbol{\Delta s}_*$, as shown in equation (1.25). The features used as split nodes in the trees are pixel-difference features, similarly to [Cao et al., 2012]. For each split node, 500 randomly sampled features are tested and the one producing maximum variance reduction is selected. After training, the leaves of the trees store 2D offset vectors that are the average of all the training samples in each leaf. The random forest thus effectively performs feature selection on the multitude of local pixel-differences.

$$\min_{w_t^l, \Phi_t^l} \sum_{i=1} \| \pi^l \circ \boldsymbol{\Delta s}_*^i - w_t^l \Phi_t^l(\boldsymbol{I}^i, \boldsymbol{s}_{t-1}^i) \|_2^2, \tag{1.25}$$

where i iterates over the training samples, operator π^l extracts $(\Delta x_*^l, \Delta y_*^l)$ from the vector $\boldsymbol{\Delta s}_*^i$, thus making $\pi^l \circ \boldsymbol{\Delta s}_*^i$ the ground truth 2D-offset of the l^{th} landmarks, in i^{th} training image. If D is the total number of leaves in the forest, w_t^l is a 2-by-D matrix, which columns store the 2D offset vector of each leaf and Φ_t^l is a D-dimensional binary vector containing ones if the test sample reaches the corresponding leaf node and zeros otherwise. Each Φ_t^l is thus a sparse

binary vector with as many non-zero elements as there are trees in the forest. These are the *local binary features* and are illustrated in figure 1.11.

Learning each local random forest results in both Φ_t^l, the feature vector, and w_t^l, the local regression. These are *discarded* and instead, all local feature vectors are concatenated into $\mathbf{\Phi}_t$, as shown in figure 1.11. A global linear regression R_t is then learned by minimizing equation (1.26).

$$\min_{R_t} \sum_{i=1}^{N} \|\Delta \boldsymbol{s}_*^i - R_t \mathbf{\Phi}_t(I^i, \boldsymbol{s}_{t-1}^i)\|_2^2. \tag{1.26}$$

Fitting Similarly to the SDM, the algorithm generates a sequence of updates, as described in algorithm 1. At each iteration, a linear regression is performed from the binary feature vector computed at the landmarks' previous locations $\boldsymbol{s}_{t-1}$ to an update $\Delta \boldsymbol{s}_t$ which is added to the current landmarks' locations estimate.

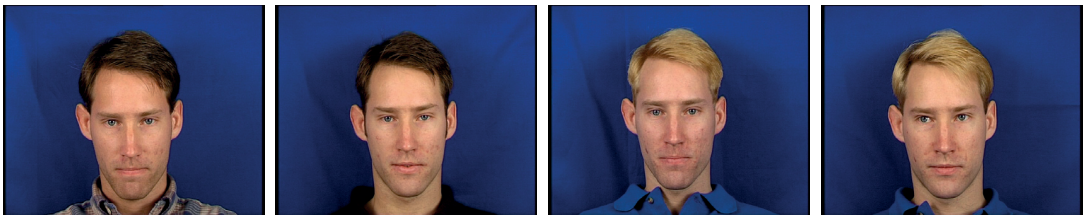
With respect to the SDM, the LBF method is much faster, as we will show in the benchmark in section 1.4. This is mainly due to the fact that the features are faster to compute.

1.4 Benchmark

In this section, the landmarks localization methods introduced in section 1.3 are benchmarked in terms of speed and accuracy. As part of this thesis, a C++ library for facial landmark localization and derived applications, *lts5-face*, was developed. The SDM and LBF have been implemented in that library, whereas the AAM[3] and CLM [4] are largely based on publicly available C++ implementations. The benchmarking framework was also implemented in the C++ library and takes advantage of the availability of the custom implementations or wrappers for each method, in order to compare them in a consistent way, in particular with respect to the timing of the different steps. In this benchmark, we use the Viola-Jones face detector to detect faces in the training and testing images. Images in which the face is not detected are discarded from the database. On the two databases used in this benchmark, the Viola-Jones face detector detects 99.8% and 97.3%, respectively.

In the next subsections, we first describe the datasets that we used in order to train and test each face model in subsection 1.4.1. Then, we present the benchmark's results and shortly discuss them in subsection 1.4.2.

[3] Active Appearance Models C++ Library, available at https://github.com/greatyao/aamlibrary
[4] CLM implementation available at https://github.com/takiyu/CLM



(a) Session 1 (b) Session 2 (c) Session 3 (d) Session 4

Figure 1.12 – Examples of frontal face images from *XM2VTS* database of subject *001* across four recording sessions.

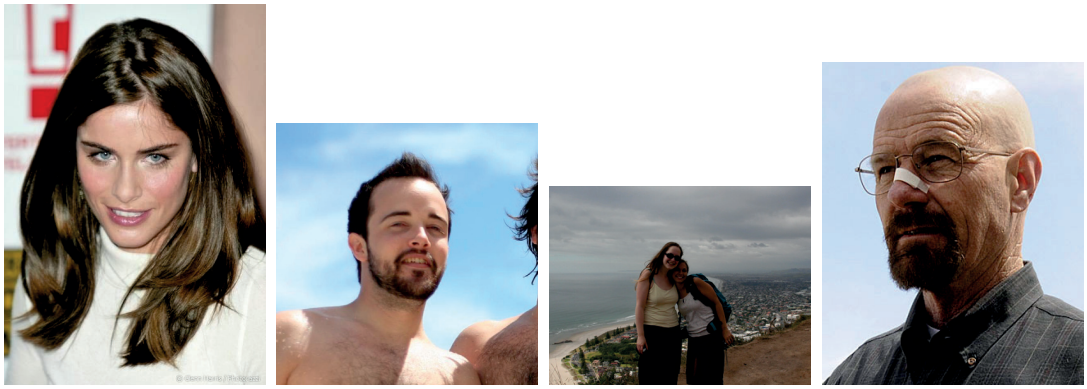
1.4.1 Datasets

In the past two decades, a certain number of databases of face images and ground-truth landmarks annotations have been proposed and used by the research community to compare facial landmark localization methods. If the images themselves are easy to acquire and massively available on the internet, for example on social networks, the same does not apply to landmarks annotations. Obtaining these still requires a lot of manual work. Moreover, these annotations result from a subjective process and can vary between different annotators. It is thus often a good practice to collect annotations from different annotators and fuse them in a way that privileges a consensus between them.

An important characteristic of a database is whether it has been captured in controlled conditions or in uncontrolled conditions, i.e. *in-the-wild*. Databases captured in controlled conditions exhibit well defined variations in term of illumination, occlusions, head pose and facial expressions. They are usually less challenging for facial landmark localization methods as they exhibit altogether less variation than *in-the-wild* databases. Conversely, *in-the-wild* databases aim at providing benchmark conditions closer to real-world scenarios, without any control on illumination, occlusions, head pose and facial expressions. Their images are often collected from publicly available sources on the internet.

In this benchmark, we present results on one database captured in controlled conditions, the *XM2VTS* database [Messer et al., 1999], and one *in-the-wild* database, *300 Faces in-the-Wild Challenge (300-W)* [Sagonas et al., 2013b, Sagonas et al., 2015]. For both of them, annotations are provided by [Sagonas et al., 2013a] and follow the Multi-PIE markup illustrated in figure 1.5.

The *XM2VTS* database [Messer et al., 1999] contains 2360 frontal images of 295 different subjects. The subjects were recorded in front of a blue background and were illuminated from both left and right sides with diffusion gel sheets to keep this illumination as uniform as possible. For each subject, eight shots were recorded during four distinct sessions over a period of four months. Figure 1.12 shows one shot of each session for subject *001*, as examples. The database was split arbitrarily into a training set and a testing set. The training set contains 1184 images of the first 148 subjects and the testing set 1176 images of the last 147 subjects.



(a) LFPW (b) Helen (c) AFW (d) IBUG

Figure 1.13 – Examples of face images from *300-W* database taken from each pre-existing databases and IBUG.

Note that all images of any given subject are either in the training set or in the testing set and no subject appears both in the training and the testing set. The Viola-Jones face detector correctly detects 99,8% of the testing set images, that is 1174 images out of 1176. The two images in which no face is detected are removed from the testing set. In both cases, the face is not detected due to an extreme pose, the head being tilted down, with the lower part of the face being out of the image.

The *300 Faces in-the-Wild Challenge (300-W)* [Sagonas et al., 2013b, Sagonas et al., 2015] database[5] was introduced for the first Automatic Facial Landmark Detection in-the-Wild Challenge held in conjunction with the International Conference on Computer Vision 2013. It is composed of face images from several pre-existing databases: Label Face Parts in the Wild (LFPW) [Belhumeur et al., 2011], containing 811 training and 224 test face images downloaded from the web on sites such as google.com, flickr.com and yahoo.com, Helen [Le et al., 2012], containing 2000 training and 330 test face images downloaded from flickr.com and Annotated Faces in the Wild (AFW) [Zhu and Ramanan, 2012], containing 337 face images downloaded from flickr.com. In addition, it contains 135 face images from a new IBUG database. We used the training sets of LFPW and Helen, as well as the complete IBUG and AFW database, as training set and the testing sets of LFPW and Helen as testing set. Thus, we use in total 3283 training images and 524 testing images. The Viola-Jones face detector correctly detects 97.3% of the faces in the testing set, that is 539 out of 554 images. The face detector fails to detect more faces, due to the unconstrained nature of the database: more images present large head-pose, bad illumination, or partial occlusion. The images in which no face is detected are removed from the testing set.

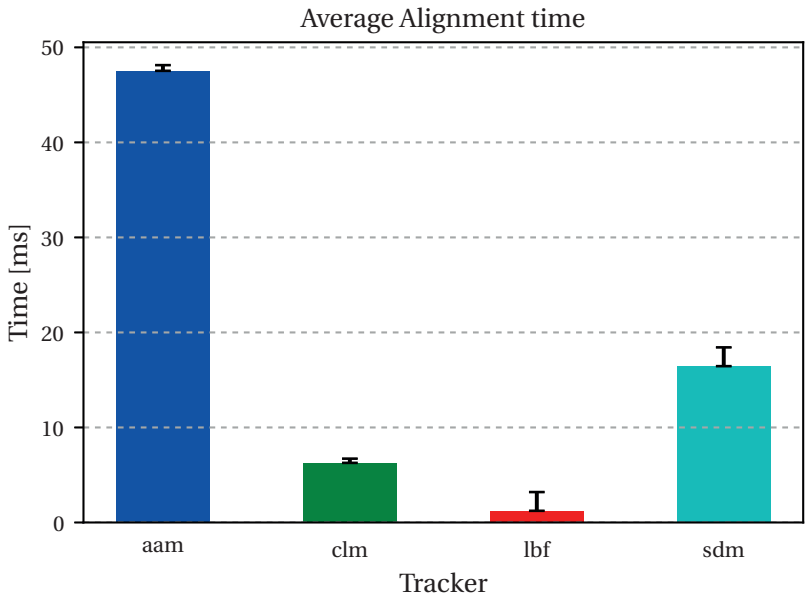


Figure 1.14 – Average face alignment time and its standard deviation for AAM, CLM, LBF, and SDM methods.

1.4.2 Results

We first compare each methods in terms of average alignment time. Given a face region from the face detector, we record the time it takes to perform facial landmark localization for each test image of the *XM2VTS* database and report the average over the 1176 images of the testing set. Note that the alignment time is independent of both the training and the testing set and solely depends on the complexity of the method. Figure 1.14 presents the results for the four methods benchmarked here: AAM, CLM, LBF and SDM.

The AAM uses the inverse compositional method and a four stages multi-resolution pyramid. It is the slowest method with the face alignment taking 48.4ms in average. This is due to the need to perform a warp of the input image at each step. The CLM is around 7 times faster, with the alignment taking 7.3ms in average, but the comparison is not completely fair as it is the only method that does not build a pyramid. Regression methods are also much faster than the AAM, with the SDM being approximately 3 times faster with the alignment taking 16.9ms in average. It uses a 4 stages pyramid. The LBF used in this benchmark is trained with a 5 stages pyramid, 5 trees per landmarks, or 340 trees in total, and a maximum depth of 5 for each tree. The LBF benefits from very simple features, which are thus fast to compute, and is approximately 10 times faster than the SDM with the alignment taking 1.8ms in average.

Then we compare the different methods in terms of accuracy. A common metric for compar-

[5]The *300-W* database is available at https://ibug.doc.ic.ac.uk/resources/300-W/

Table 1.1 – Results on the XM2VTS scenario

Method	RMSE	CED(0.05)	CED(0.1)	CED(0.2)
AAM	0.1240	0.0051	0.3134	0.9421
CLM	0.0864	0.1311	0.7521	0.9719
LBF	0.0452	0.7402	0.9642	0.9957
SDM	**0.0375**	**0.8577**	**0.9931**	**0.9991**

ing facial landmark localization methods is the root mean square error (RMSE) between the ground truth and the computed facial landmarks over the complete set of facial landmarks, normalized by the inter-ocular distance. This metric is referred to as the normalized RMSE and the smaller its value, the better. The normalized RMSE is informative to compare the overall performances of different methods on a given database, but it does not provide any information about the distribution of the errors across the landmarks. Often, landmarks which define external contours of the face, for example those along the chin, are less accurately localized than landmarks within the face, such as those around the eyes or the mouth. Normalized RMSE also does not provide information about the distribution of the errors across images. Moreover, it is negatively influenced by even a small number of outliers.

In order to get a more detailed understanding of the distribution of the errors across images, the cumulative error distribution (CED) is very often used. It represents the percentage of the images in the database for which the error is smaller than a given value. We also report values of the CED for different errors level: 0.05, 0.1, and 0.2. These values indicate the percentage of the test images for which the normalized RMSE is smaller than 5%, 10%, and 20% of the inter-ocular distance, respectively. In [Dantone et al., 2012] and [Burgos-Artizzu et al., 2013], the authors consider that a normalized error higher than 0.1 is a failure. Even though that threshold might be rather conservative, it gives an intuitive understanding of what can be considered a good performance. These metrics are computed and compared across three different test scenarios: the **XM2VTS** scenario, the **300-W** scenario and the **cross-databases** scenario.

XM2VTS results

In this scenario, each model is trained on *XM2VTS* training set and tested on *XM2VTS* testing set. This corresponds to the less challenging scenario, as both the training and testing set exhibit limited variation. Moreover, the training set is representative of the testing set, as both are part of the same database.

Table 1.1 presents the RMSE, and the three values of the CED obtained on the *XM2VTS* database and figure 1.15 presents the corresponding CED. A few examples of alignments from the testing set of the *XM2VTS* database are shown in figure 1.16 for each of the methods.

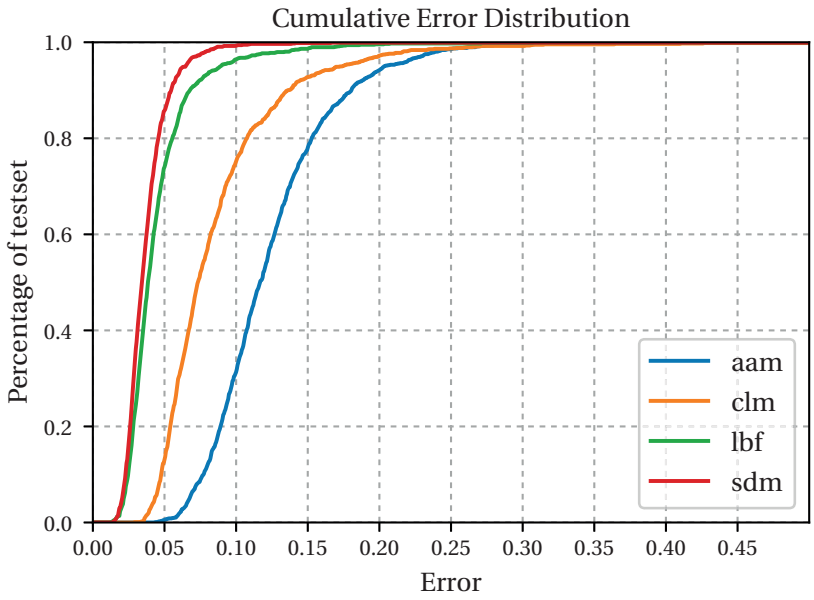


Figure 1.15 – Cumulative error distribution on the *XM2VTS* database.

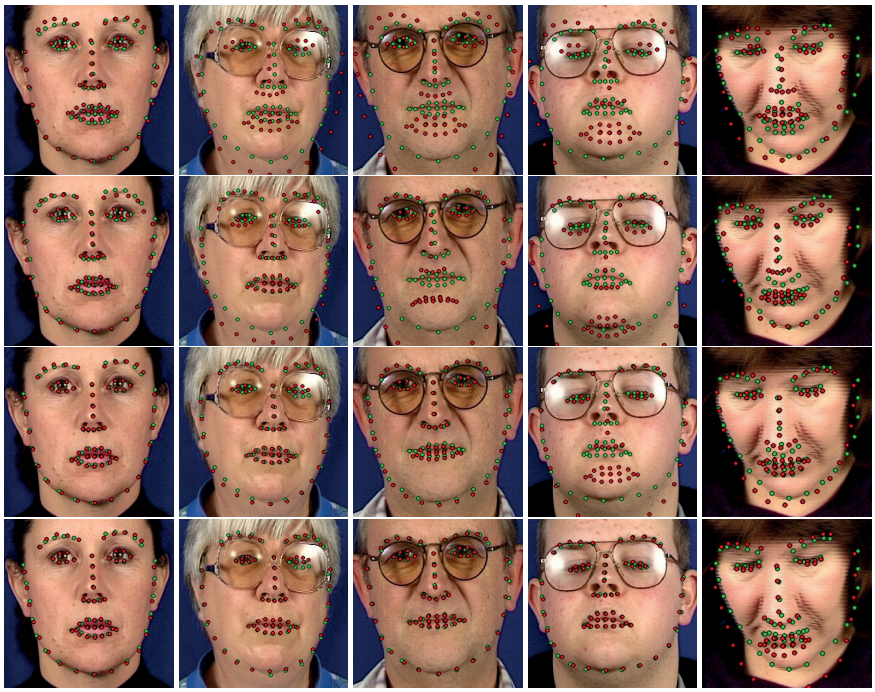


Figure 1.16 – Examples of fits on the *XM2VTS* database with the AAM (first row), the CLM (second row), the LBF (third row) and the SDM (last row). The annotations are in green (●) and the resulting facial landmarks are superimposed in red (●). From left to right, simpler to more challenging images. Every methods succeed on the left most image and fail on the right most image.

Table 1.2 – Results on the 300-W scenario

Method	RMSE	CED(0.05)	CED(0.1)	CED(0.2)
AAM	0.8571	0.0	0.0	0.0018
CLM	0.1667	0.0055	0.1892	0.7514
LBF	0.0838	0.1725	0.8071	0.9666
SDM	**0.0756**	**0.2338**	**0.8757**	**0.9907**

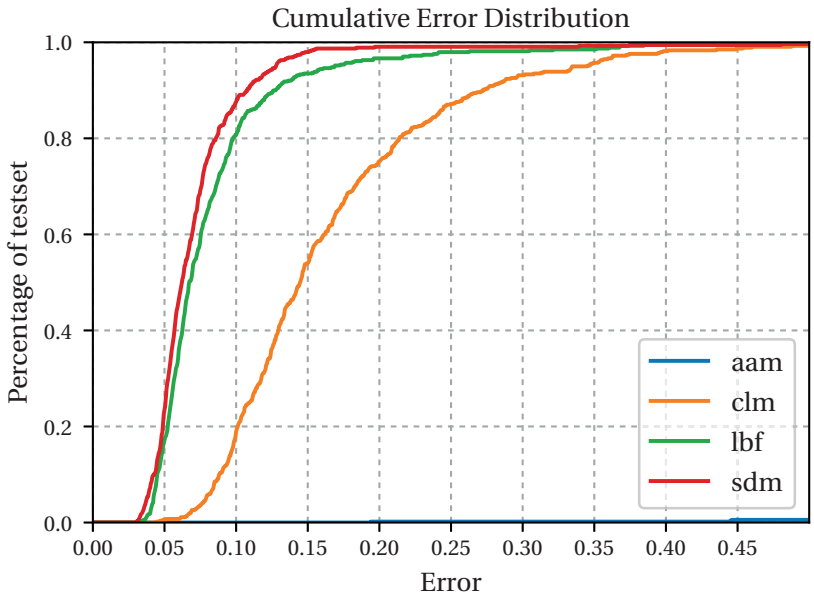


Figure 1.17 – Cumulative error distribution on the *300-W* database

Even on this database, captured in controlled conditions, the AAM results in an RMSE approximately three times higher than the SDM and LBF, the two regression-based methods. It shows that the AAM's performances on previously unseen subjects, even on mostly frontal images, without facial expressions or occlusions, are limited. Only 31.34% of the testing set has a normalized error lower than 0.1. On the other hand, regression-based methods perform very well, 96.42% and 99.31% of the faces in the testing set are fitted with a normalized error smaller than 0.1, using LBF and SDM, respectively. The RMSE of the LBF is only 0.77% above the RMSE of the SDM. These two methods are thus very similar in terms of performances while the LBF is approximately 10 times faster than the SDM. This represents a real advantage in real-time tracking applications.

300-W results

In this second scenario, each model is trained on *300-W* training set and tested on *300-W* testing set. As detailed in section 1.4.1, *300-W* is an *in-the-wild* database and as such is much

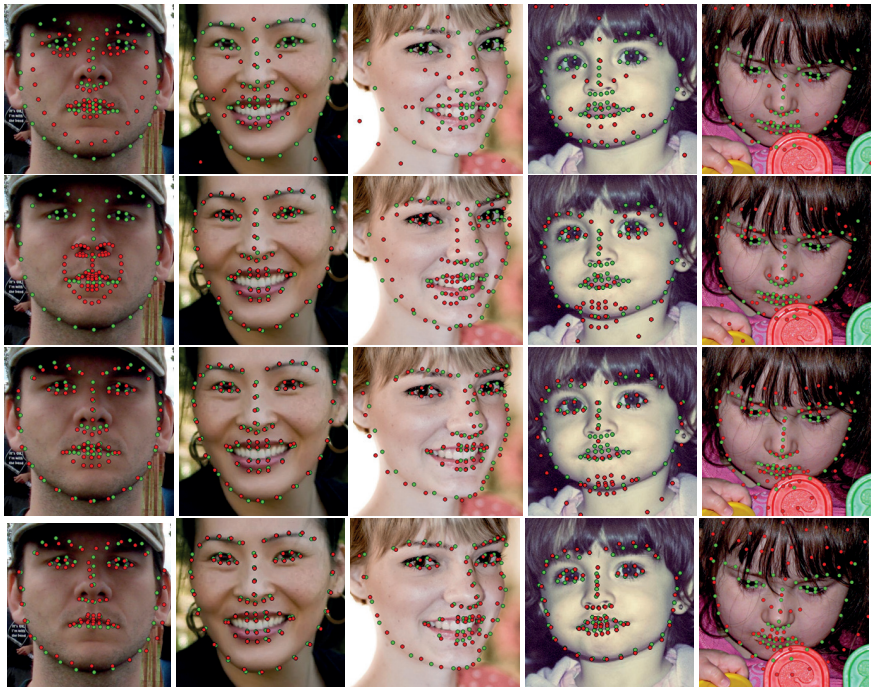


Figure 1.18 – Examples of fits on the *300-W* database with the AAM (first row), the CLM (second row), the LBF (third row) and the SDM (last row). The annotations are in green (●) and the resulting facial landmarks are superimposed in red (●).

more challenging than *XM2VTS*.

Table 1.2 presents the RMSE, and the three values of the CED obtained on the *300-W* database and figure 1.17 presents the corresponding CED. A few examples of alignments from the testing set of the *300-W* database are shown in figure 1.18 for each of the methods.

On this challenging, *in-the-wild* database, the AAM performances are extremely bad, showing the AAM's incapacity to learn a proper representation of a face on a training set with as much variation. The RMSE of the CLM has doubled with respect to the first scenario on the *XM2VTS* database and only 18.92% of the faces in the testing set are fitted with a normalized error smaller than 0.1. The two regression-based methods still perform well and the difference between them remains close to 0.8%.

Cross-databases results

In the third scenario, each model trained on *XM2VTS* training set is tested on *300-W* testing set. This is the most challenging scenario. The training set exhibits limited variations in terms of head pose, facial expressions, and illumination. The testing set, on the other hand, exhibits large variations with respect to these factors.

Table 1.3 presents the RMSE, and the three values of the CED obtained on the *300-W* database

Table 1.3 – Results on the cross-database scenario

Method	RMSE	CED(0.05)	CED(0.1)	CED(0.2)
AAM	0.2126	0.0	0.0130	0.5306
CLM	0.1812	0.0	0.1336	0.6698
LBF	0.1653	0.0019	0.1837	0.7570
SDM	**0.1162**	**0.0186**	**0.5492**	**0.9165**

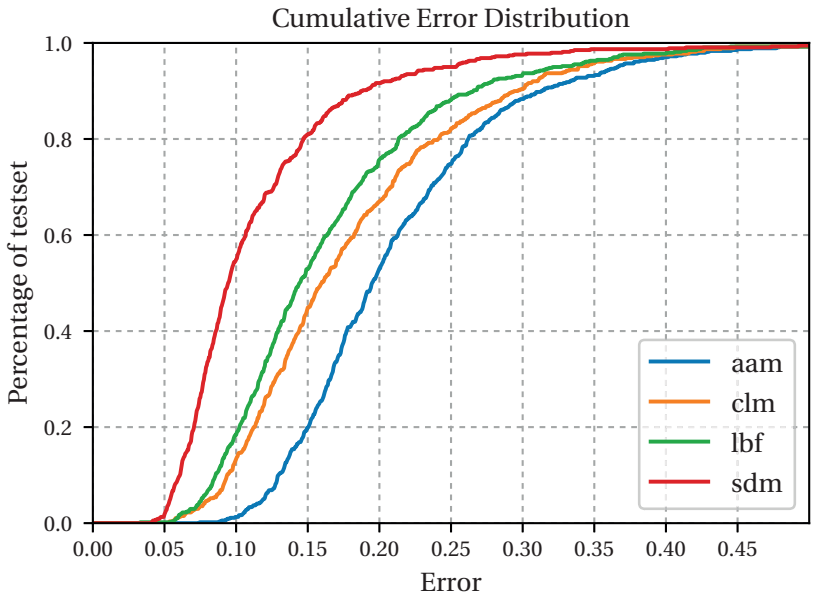


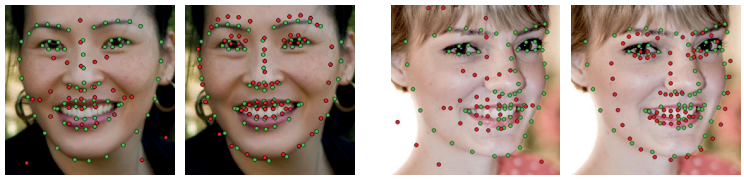
Figure 1.19 – Cumulative error distribution on the cross-database scenario

and figure 1.19 presents the corresponding CED.

Except for the AAM, the performances decrease with respect to the second scenario, which is using the same testing set but a different training set. In contrast to the *300-W* scenario, the AAM did learn a representation of the face from the more constrained training set of *XM2VTS*. Figure 1.20 shows a comparison between alignments obtained on the testing set of the *300-W* database when training the model on the same database or on the more consistent training set of the *XM2VTS* database. Nevertheless, all methods except the AAM suffer from the fact that the training set is not representative of the testing set in terms of variation, since it is recorded in constrained conditions.

1.5 Conclusion

In this chapter, we have described a typical facial image analysis pipeline and representative methods used in the face acquisition step. A face detector first detects faces in the image and



(a) (b) (c) (d)

Figure 1.20 – AAM improvement when trained on the *XM2VTS* training set. The annotations are in green (●) and the resulting facial landmarks are superimposed in red (●). (a) and (c) Results on an image from the *300-W* testing set when trained on the *300-W* training set. (b) and (d) Results on an image from the *300-W* testing set when trained on the more consistent *XM2VTS* training set.

returns the corresponding bounding boxes. In section 1.2, we have described two standard methods for face detection: the Viola-Jones face detector and Yang and Ramanan's parts-based detector. The Viola-Jones face detector, the first real-time face detector, is fast but its performances are hindered by its holistic representation of the face. This makes it less suitable to detect faces in a collection of images containing a lot of variation in terms of head pose, facial expressions, or occlusions. Conversely, Yang and Ramanan's local parts-based detector overcomes the limitations of a holistic model but is also slower, thus not being adapted to real-time performances.

From the bounding box of a face, a plethora of methods have been proposed to perform facial landmark localization. In section 1.3, we have presented a categorization of these methods into four categories and a timeline of their development, spanning more than 20 years. We have also described four of these methods, from three different categories: the AAM, a simple CLM, the SDM and the LBF.

Finally, in section 1.4 we have proposed a benchmark of these four methods on two different publicly available databases, the *XM2VTS* database, recorded in controlled conditions, and the *300-W* database, an *in-the-wild* database. In our benchmark, the SDM performs consistently better than the other methods and requires around 17ms in average to localize landmarks on a face. In terms of performance, the LBF is very close to the SDM on both databases and presents the additional advantage of being almost 10 times faster, with only 1.8ms in average per face. This makes it very suitable for real-time applications or applications on mobile platforms.

In general, we have tried to give an overview of the core methods in facial image analysis. We hope that this chapter can provide a smooth introduction to this field and help the newcomer to understand its development and the remaining challenges.

Part I

2D facial image analysis for automatic prediction of difficult intubation

Overview

In this first part, we focus on difficult intubation prediction, a medical diagnosis problem in anesthesiology. In chapter 2, we first introduce the topic by reviewing some definitions and existing methods of prediction of the difficult tracheal intubation and discuss their limitations. This chapter aims at providing a basic understanding of the difficult tracheal intubation prediction problem, from a medical point-of-view, to the reader without a medical background.

Chapter 3 presents a method to classify images of patients, with the mouth wide open and the tongue protruding to its maximum, according to their modified Mallampati score, a simple indicator of potential difficulty to intubate, described in chapter 2. This method is trained and tested on 100 patients annotated by experienced anesthesiologists. We first extract appearance based features, derived from the active appearance model (AAM) shape-free appearance component, then perform feature selection with a linear support vector machine (SVM), and finally classify each image into one of the four modified Mallampati score. Our system achieves a high accuracy of 95% in a leave-one-subject-out cross validation scheme. Even though the clinical value of the modified Mallampati score is criticized, when this test is used alone, it is often used in most of the multifactorial tests. As such, the results obtained in this chapter can be considered as encouraging preliminary results for integration of the modified Mallampati classification into a more complete, fully automated method, which would consider other factors as well.

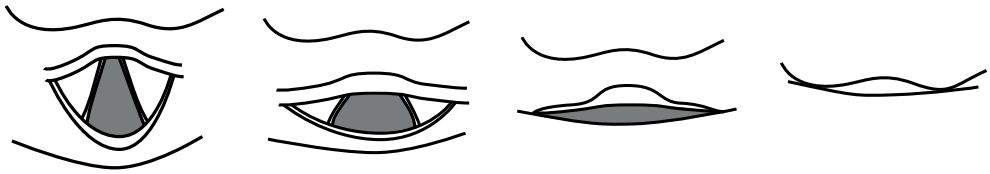
In the final chapter of this initial part, chapter 4, we present a completely automatic method, based on facial morphometry and extending our work on modified Mallampati, to predict a patient's difficulty of intubation with performance comparable to medical diagnosis-based predictions by experienced anesthesiologists. We also give insights on the possible limitations of the method and comment on the utility of a three dimensional (3D) face model and analysis methods, as presented in part II.

The different contributions of this part have been published in [Cuendet et al., 2012] and [Cuendet et al., 2015]. A patent is also pending for the method described in chapter 4 [Schoettker et al., 2014].

2 Introduction to the prediction of difficult tracheal intubation

The priority of the anesthesiologist, after having induced general anesthesia is to ventilate the patient and secure his airways. As the patient is under the influence of drugs, whose main effects are the loss of consciousness, analgesia, and muscular paralysis, he is unable to breath by himself and mechanical ventilation is mandatory. Despite all the advancements in anesthesiology, difficult airway management still represents a major cause of anesthesia-related injuries with potential life threatening complications [Peterson et al., 2005]. Recent analysis of airway management related claims in the UK [Cook and Macdougall-Davis, 2012] and in the USA [Metzner et al., 2011] show that respiratory events, most of them being difficult intubation or inadequate ventilation, come first in the proportion of cases with poor clinical outcomes, ranging from severe harm to brain damage or death. The worst case scenario in airway management is the *"Can't intubate, can't ventilate"* situation, in which the patient is impossible to be ventilated by face mask and intubated with an endotracheal tube. The estimated incidence of such a situation is estimated between 0.01 and 3 in 10'000 cases [Heard et al., 2009]. Nowadays, up to one third of all deaths attributed to anesthesia are consecutive to the inability to either ventilate or intubate [Hove et al., 2007]. Numerous technical advances have allowed facilitation of intubation by improving the view at laryngoscopy [Aziz et al., 2011, Teoh et al., 2010, Serocki et al., 2010] or monitoring the placement of the endotracheal tube [Juan et al., 2002, Räsänen et al., 2006]. Yet, difficult intubation still remains an area of concern [Cook and Macdougall-Davis, 2012, Hung et al., 2016].

Detection and anticipation of difficult airway in the preoperative period is crucial for patients' safety. In cases of suspected difficulty, specific equipment and personnel will be called upon to increase safety and the chances of successful intubation. In daily practice, anesthesiologists predict the difficulty of tracheal intubation with bedside tests, which correlate poorly with the ground truth. Experienced anesthesiologists associate, in addition to the available bedside tests, a global clinical judgment, probably based on a larger number of morphological parameters than those contained in the available bedside tests described in this chapter. Nevertheless a high proportion of patients with a difficult airway remain undetected despite the most careful preoperative airway evaluation. According to the Danish Anaesthesia Database



(a) Grade 1 (b) Grade 2 (c) Grade 3 (d) Grade 4

Figure 2.1 – Four grades of the Cormack-Lehane classification of the laryngoscopic view.

[Nørskov et al., 2015], which included 188,064 patients, the diagnostic accuracy of the anesthesiologists' predictions of difficult laryngoscopic intubation and difficult mask ventilation was poor. Specifically, out of 3391 difficult intubations, 3154 (93%) were unanticipated and out of 857 cases of difficult mask ventilation, 808 (94%) were unanticipated.

In this chapter, we aim to provide the necessary background to the reader without a medical training, in order to understand the difficult tracheal intubation prediction problem. We first review the definitions of the difficult tracheal intubation in section 2.1. A major problem when defining the difficulty of tracheal intubation is the inherent variability in which the difficulty is evaluated: different conditions, at different moments, and with different anesthetists influence the difficulty of tracheal intubation. In section 2.2, we then review existing methods and bedside tests for the prediction of difficult tracheal intubation. We then summarize and conclude this chapter in section 2.3.

2.1 Definitions of the difficult tracheal intubation

For the last 30 years, numerous definitions have been proposed and used by anesthesiologists, but no unique definition of difficult intubation exists. The vast majority of endotracheal intubations are performed using a laryngoscope which allows the visualization of the larynx and the placing of the endotracheal tube between the vocal cords, into the trachea. Thus, one of the first attempt to define difficult intubation objectively was by associating a difficult intubation with a difficult laryngoscopy.

2.1.1 Cormack-Lehane classification of the laryngoscopic view

Cormack and Lehane proposed a classification of the laryngoscopic view using four grades based on the visibility of laryngeal structures or glottic exposure [Cormack and Lehane, 1984]. Figure 2.1 illustrates the view of the different anatomical structures for each grade of the Cormack-Lehane classification.

This classification was later modified by Yentis and Lee who proposed to divide the original grade 2 into grade 2a and grade 2b [Yentis and Lee, 1998]. The later classification is used to

define the difficult laryngoscopy as a view corresponding to grade 3 or grade 4. Nevertheless, it has recently been pointed out by Krage et al. that the reproducibility of this classification is limited [Krage et al., 2010]. Moreover, a poor view of the vocal cords can increase the difficulty of the intubation but other factors, such as the position of the head of the patient or the experience of the anesthesiologist also have influence on the success of the intubation.

Various national societies of anesthesiology have set their own definitions of difficult intubation. In France, the *Société Française d'Anesthésie et Réanimation (SFAR)* qualifies an intubation as difficult *"When more than two laryngoscopies are performed and/or an alternative technique is used after head position optimization, with or without external laryngeal manipulation"* [Diemunsch et al., 2008]. In the USA, the *American Society of Anesthetists (ASA)* says of an intubation that it is difficult *" when tracheal intubation requires multiple attempts, in the presence or absence of tracheal pathology"* [Caplan et al., 2003, Apfelbaum et al., 2013].

Despite the need for a standard classification of the difficult intubation in the medical community, no such uniform definition has been widely adopted. Thus, the incidence and the factors associated with difficult intubation vary from one institution to another and are virtually impossible to compare directly. The incidence of difficult laryngoscopy in the operating room has been reported to range from 0.3% to 13% [Naguib et al., 1999].

2.1.2 Adnet's Intubation Difficulty Scale

In an attempt to provide a definition of the difficult intubation, Adnet et al. proposed the *intubation difficulty scale (IDS)* [Adnet et al., 1997], taking into account the number of attempts, the number of operators directly attempting the intubation, the use of alternative devices or techniques, the glottic exposure or the lifting force applied during laryngoscopy.

Their hypothesis is that what characterizes the difficulty of an intubation is how much it deviates from an *ideal* intubation performed without effort, on the first attempt, with full visualization of the laryngeal aperture and vocal cords abducted. Such an *ideal* intubation would get a score of 0. The more the intubation deviates from that situation, the more the score increases, as shown in table 2.1. Thus, the IDS is a quantitative measure of the difficulty of a specific intubation act of a patient. Nevertheless, there are no guarantees that the same patient would get the same IDS score when intubated by a different anesthetist in different conditions. The difficulty of intubation associated with each IDS score is given in table 2.2.

2.2 Methods of prediction of the difficult tracheal intubation

Prediction of difficult endotracheal intubation has been largely explored in the past twenty-five years by anesthesiologists. Several physical and morphological characteristics have been identified as predictors of difficult laryngoscopy or difficult intubation. Those include: obesity, poor mobility of the head and neck, poor mobility of the jaw, receding mandible, long upper

Table 2.1 – Intubation Difficulty Scale [Adnet et al., 1997]

Parameter	Rule	
Number of attempts >1	Every additional attempt adds 1 pt	N_1
Number of operators >1	Every additional operator adds 1 pt	N_2
Number of alternative techniques	Each alternative technique adds 1 pt	N_3
Cormack grade - 1	Apply Cormack grade for first oral attempt. For successful blind intubation $N_4 = 0$	N_4
Lifting force required		
Normal		$N_5 = 0$
Increased		$N_5 = 1$
Laryngeal pressure	Sellick's maneuver adds no points	
Not applied		$N_6 = 0$
Applied		$N_6 = 1$
Vocal chords mobility		
Abduction		$N_7 = 0$
Adduction		$N_7 = 1$
IDS		$\sum_{i=1}^{7} N_i$

Table 2.2 – Degree of difficulty given the IDS score [Adnet et al., 1997]

Score	Degree of difficulty
0	Easy intubation
$0 < \text{IDS} \leq 5$	Slight difficulty
$5 < \text{IDS}$	Moderate to Major difficulty
$\text{IDS} = \infty$	Impossible intubation

incisors, decreased mouth opening, or small interincisor gap with the mouth fully open, shortened thyromental distance (TMD), short neck and small neck circumference. Several difficult intubation bedside screening tests exist.

2.2.1 Patil-Aldreti test, or thyromental distance

The thyromental distance (TMD), or Patil-Aldreti test, is the distance from the upper edge of the thyroid cartilage to the chin, measured with the head fully extended. A short thyromental distance equates to an anterior lying larynx that is at a more acute angle and also results in less space for the tongue to be compressed by the laryngoscope blade. A thyromental distance greater than 7 cm is usually associated with easy intubation whereas a thyromental distance smaller than 6 cm may predict a difficult intubation.

However, with a sensitivity of 48% and a specificity of 79% in predicting difficult intubation [Baker et al., 2009], this distance is not a good predictor by itself and is often used in combination with other predictors. The ratio of height to thyromental distance (RHTMD) improves the accuracy of predicting difficult laryngoscopy compared to TMD alone with a sensitivity and specificity of 77% and 54% respectively [Krobbuaban et al., 2005].

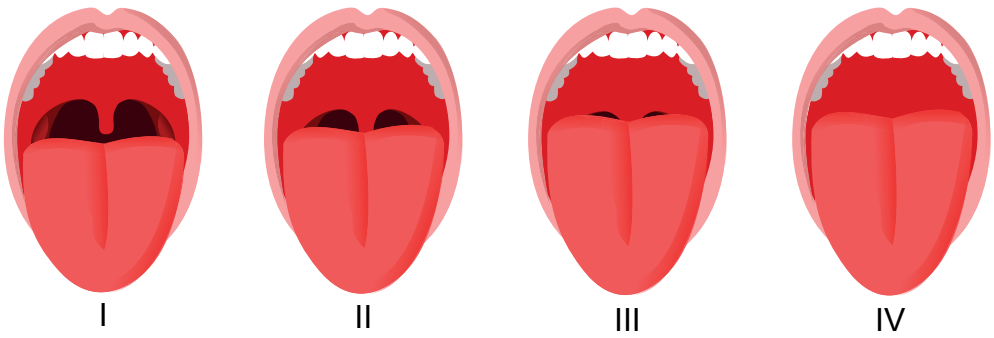


Figure 2.2 – The four grades of the Mallampati score
(source: Wikimedia Commons, author: Jordi March i Nogué, CC-BY-SA 3.0)

2.2.2 Mallampati score

Originally described by Mallampati et al. [Mallampati et al., 1985] and modified by Samsoon and Young [Samsoon and Young, 1987], the Mallampati score assesses the airway according to the visibility of oropharyngeal structures observed on a sitting patient with the mouth wide open and the tongue out. The hypothesis of the author is that the larger the base of the tongue, the more it overshadows the larynx, resulting in a poor laryngoscopic view and a potentially difficult laryngoscopy. The volume of the tongue is thus an important, yet difficult to assess, parameter when assessing the difficulty of endotracheal intubation. Since it is not possible to determine the volume of the tongue relative to the capacity of the oropharyngeal cavity, it is logical to infer that the base of tongue is disproportionately large when it is able to mask the visibility of the faucial pillars and uvula.

The score ranges from class 1 to class 4. Class 1 indicates full visibility of the oropharyngeal structure: the soft palate, fauces, uvula, and pillars are visible. Class 2 indicates a reduced visibility: only soft palate, fauces, and uvula are visible. Class 3 indicates a limited visibility: the soft palate and only the base of the uvula are visible. Class 4 indicates no visibility: the soft palate is not visible at all. Figure 2.2 illustrates the four grades of the Mallampati score.

Various meta-analysis reported different sensitivity and specificity for the Mallampati and modified Mallampati tests. In [Cattano et al., 2004], the authors reported a sensitivity and a specificity of 35% and 91% respectively. In [Lundstrøm et al., 2011], the authors included 55 studies and 177088 patients and reported a sensitivity of 0% to 100% and a specificity of 44% to 100%. They computed a receiver operating characteristic (ROC) curve and the area under the curve (AUC) was 0.753 which categorize the diagnostic test as good. In [Lee et al., 2006], the reported AUC for the Mallampati and modified Mallampati tests are respectively 0.58 and 0.83. In those studies, the authors agree that the clinical value of the Mallampati test is limited as it has poor to moderate discriminative power when used alone.

Table 2.3 – Wilson Risk Sum Score [Wilson et al., 1988]. IG = interincisor gap; SLux = subluxation (maximal forward protrusion of the lower incisors beyond the upper incisors)

Risk factor	Level	Variable
Weight	0	< 90kg
	1	90 – 110kg
	2	> 110kg
Head and neck movements	0	> 90°
	1	About 90° (i.e., ±10°)
	2	> 90°
Jaw movement	0	IG ≥ 5cm or SLux > 0
	1	IG < 5cm and SLux = 0
	2	IG < 5cm and SLux < 0
Receding mandible	0	Normal
	1	Moderate
	2	Severe
Buck teeth	0	Normal
	1	Moderate
	2	Severe

2.2.3 Upper lip bite test

The upper lip bite test, proposed by Khan et al. [Khan et al., 2003] evaluates the ability of the patient to cover his upper lip with the lower incisors by moving forward the lower jaw in a movement of *prognathism*. The results range from grade I to grade III where grade I and II predicts easy laryngoscopy whereas grade III predicts difficult laryngoscopy. The authors initially observed a sensitivity of 76.5% and a specificity of 88.7%. Those results were confirmed in a recent study in which the authors obtained 78.95% and 91.96% respectively [Khan et al., 2009].

Eberhart et al. conducted a comparison between Mallampati score and upper lip bite test on 1107 patients [Eberhart et al., 2005] and concluded that both tests are poor predictors for difficult laryngoscopy when used as single preoperative bedside screening tests.

None of those simple tests have been shown to be accurate in predicting airway management problems. Their sensitivity and predictive positive values are generally low, precluding an accurate prediction of difficult endotracheal intubation. Thus, several studies have been proposed to derive a score from multivariate analysis. The three most common multivariate bedside screening tests are the Wilson risk sum score, the Arné model and the Naguib model and are detailed here after.

2.2.4 Wilson risk sum score

The Wilson risk sum score [Wilson et al., 1988] scores five of the aforementioned factors from 0 to 2: the weight, the vertical head and neck movement, the jaw movement, or prognathism, the

Table 2.4 – Arné simplified score model [Arné et al., 1998]

Risk factor	Score
Previous knowledge of difficult intubation	
No	0
Yes	10
Diseases associated with difficult intubation	
No	0
Yes	5
Clinical symptoms of airway pathology	
No	0
Yes	3
IG and mandible subluxation	
IG ≥ 5cm or SLux > 0	0
3.5cm < IG < 5cm and SLux = 0	3
IG < 3.5cm and SLux < 0	13
Thyromental distance	
≥ 6.5cm	0
< 6.5cm	4
Maximum range of head and neck movement	
more than 100°	0
About 90° (±10°)	2
less than 80°	5
Mallampati score	
class 1	0
class 2	2
class 3	6
class 4	8
Total possible	48

receding mandible and buck teeth as detailed in table 2.3. By varying the threshold values on the sum of those scores, the true positive rate and false positive rate of difficult laryngoscopy assessment are varied. The authors initially proposed a threshold value of 4, i.e. a score greater or equal to 4 predicts a difficult endotracheal intubation. In [Shiga et al., 2005] the authors compiled a meta-analysis of 5 studies including 6076 patients with a threshold value of 2 and reported a pooled sensitivity of 46% (95% CI, 36–56) and a pooled specificity of 89% (95% CI, 85–92). In [Naguib et al., 2006], with a thresold value of 4, the authors reported a sensitivity of 40.2% (95% CI, 30.0–50.0) and a specificity of 92.8% (95% CI, 88.0–98.0).

2.2.5 Arné model

Arné et al. proposed a simplified score model [Arné et al., 1998]. In addition to the morphological criteria such as interincisor gap, ability to prognate, thyromental distance and range of head and neck movement, it also considers the medical history of the patient and

the Mallampati score, as detailed in table 2.4. In [Arné et al., 1998] the authors computed the optimal threshold using a ROC curve. They reported an AUC of 95.6% and a sensitivity of 93% (95% CI, 80.1-98.5) and specificity of 93% (95% CI, 91.4-94.5).

2.2.6 Naguib models

Naguib et al. performed a clinical, radiologic and 3D computer imaging study [Naguib et al., 1999] on 57 patients among which 25 had an unanticipated difficult intubation. A multivariate discriminant analysis was performed on the clinical measurements and identified four risk factors that correlated with the difficult laryngoscopy and intubation: thyrosternal distance (TSD), thyromental distance (TMD), neck circumference (NC) and Mallampati classification. They proposed the following discriminant function based on these clinical criteria only:

$$l = 4.9504 + 1.1003 \cdot \text{TSD} - 2.6076 \cdot \text{Mallampati} + 0.9684 \cdot \text{TMD} - 0.3966 \cdot \text{NC}.$$

In [Naguib et al., 2006] Naguib et al. introduced a new logistic regression analysis and identified four risk factors correlated with difficult laryngoscopy and intubation: the TMD, the interincisor gap (IG), the height and the Mallampati score. They proposed the following discriminant function:

$$l = 0.2262 - 0.4621 \cdot \text{TMD} + 2.5516 \cdot \text{Mallampati} - 1.1461 \cdot \text{IG} + 0.0433 \cdot \text{height}.$$

The authors reported an AUC of 90% when tested on 194 patients. In [Langeron et al., 2012] the authors report an AUC of 66% for the same test conducted on 1655 patients among which 101 (6.10%) were difficult to intubate.

2.2.7 Comparison of multivariate models and other tests

Table 2.5 shows the predictive performance of those four multivariate models, as reported in [Naguib et al., 2006]. Figure 2.3 shows the corresponding ROC curves. The authors recruited 194 patients (97 with a difficult airway and 97 controls) over a period of 5 years. For the purpose of their study, unanticipated difficult intubation was defined as difficult laryngoscopy, corresponding to a grade 3 or 4 Cormack and Lehane laryngoscopic view, and difficult tracheal intubation, with 2 or more attempts at placing the endotracheal tube, or the use of an alternative device, such as laryngeal mask airway or bougie, when using optimal head and neck positioning (the sniffing position). Positive predictive value (PPV) and negative predictive value (NPV) were calculated based on a prevalence of difficult intubation of 5.8%, as reported in a recent meta-analysis [Shiga et al., 2005]. Note that the sensitivity, specificity and AUC are the most appropriate measures to compare performance between datasets, mainly due to the class imbalance problem.

Table 2.5 – Comparison of four multivariate tests [Naguib et al., 2006]

Model	Sens.	Spec.	PPV	NPV	AUC	Acc.
Wilson model[Wilson et al., 1988]	40.2	92.8	25.6	96.2	79.0	66.5
Arné model [Arné et al., 1998]	54.6	94.9	39.7	97.1	87.0	74.7
Naguib model I [Naguib et al., 1999]	81.4	72.2	15.3	98.4	82.0	76.8
Naguib model II[Naguib et al., 2006]	82.5	85.6	26.1	98.8	90.0	84.0

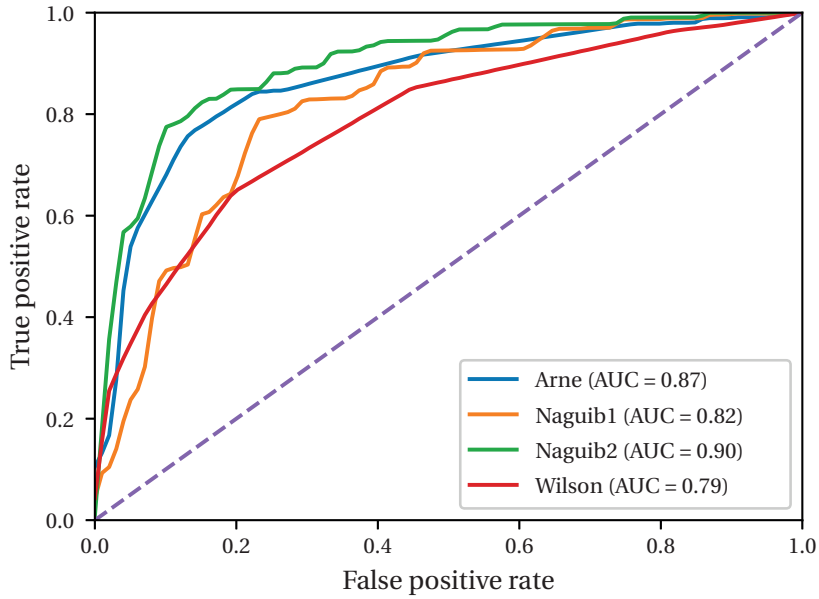


Figure 2.3 – Comparison of the ROC curves of four multivariate tests [Naguib et al., 2006]

Recently, Fritscherova et al. [Fritscherova et al., 2011] conducted a case-control study on 148 patients and concluded that the three statistically higher predictors were the interincisors distance, the TMD and a decreased temporomandibular joint movement.

As none of those tests fulfill the high sensitivity and high positive predictive value criteria, anesthesiologists themselves do not agree on the usefulness of such a prediction [Yentis, 2002, Orozco-Díaz et al., 2010].

New technological approaches aimed at craniofacial phenotyping, using still photographs, x-ray technologies or laser scanning with an automated three-dimensional rendering, have been recently applied to the detection of difficult airways.

Suzuki et al. calculated five ratios and angles from measurements derived from placement of anatomic markers on patients' photographs [Suzuki et al., 2007] demonstrating that the submandibular angle seemed to be associated with difficult tracheal intubation. They also used morphing software to construct "average" easy and difficult to intubate faces.

The improved availability of cone-beam computed tomography, 3D imaging and computer simulation has been used by Schendel and Hatcher for evaluation of the airway [Schendel and Hatcher, 2010]. In the recent years, some studies took advantage of machine learning [Langeron et al., 2012] or statistical face models [Connor and Segal, 2011] in order to provide better prediction and defend the usefulness of preoperative difficult tracheal intubation prediction. However, these newer methods require either x-ray or computed tomographic imaging methods with issues such as availability, cost and radiation dose to the patient. More recently, Cattano et al. proposed a new assessment form on airway prediction but showed that it did not improve resident ability to predict a difficult airway [Cattano et al., 2013].

Finally, the number of patients considered to validate those newer approaches is often low. For instance, in [Connor and Segal, 2011] the authors reported results on a validation set of only 20 difficult and 20 easy patients thus not demonstrating the generalizability of the proposed method. In comparison, our proposed method, described in chapter 4 has been developed and validated using more than nine hundred patients.

2.3 Conclusion

In this chapter, we first emphasized the importance of the detection and anticipation of difficult airway for patients' safety. Despite numerous technical advances, difficult intubation still remains an area of concern, according to recent studies.

Different definitions of difficult tracheal intubation are also introduced in this chapter. The diversity in these definitions and lack of reproducibility, due to the influence of the conditions and personnel, make it difficult to compare the incidence and the factors associated with difficult tracheal intubation between institutions. Nevertheless, Adnet's intubation difficulty scale (IDS) takes into account several important indicators, showing a deviation from an ideal intubation, and can be considered as a standard for quantifying the difficulty of a tracheal intubation.

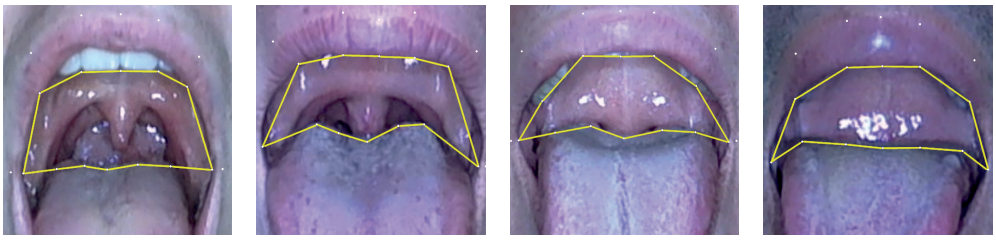
We also reviewed some of the most common bedside tests and showed that, as none of those tests fulfill the high sensitivity and high positive predictive value criteria, anesthesiologists themselves do not agree on their usefulness. Moreover, most of these bedside tests are based on morphological characteristics, which are difficult to extract in an objective, reproducible way, even by trained anesthesiologists. Based on that assessment, we hypothesize that an automatic method, based on facial morphometry, could fill the gap by extracting morphological features in an objective way and learning which features are the most discriminative, in terms of difficult tracheal intubation prediction.

3 Automatic Mallampati classification

3.1 Introduction

Assessment of difficult tracheal intubation prior to anesthesia induction is an important research topic in anesthesia and several screening tests have been proposed, detailed in section 2.2. Among them, the modified Mallampati score [Samsoon and Young, 1987] is commonly used by anesthesiologists to predict the difficulty of intubation. This score classifies the airway into 4 classes according to the visibility of the oro-pharyngeal structures observed on a patient opening the mouth and sticking his tongue out. Figure 3.1 shows one real example of each class and can be compared with the schematic representation of figure 2.2.

Although it has been shown to have little discriminative power in predicting tracheal intubation difficulty when used alone, the modified Mallampati test is still an important source of information when used in combination with other measures [Lundstrøm et al., 2011]. Among the various commonly used predictive models of difficult intubation, which use the modified Mallampati test, lies the Arné model where a simplified score is computed depending on certain physiological factors and the medical history of the patient [Arné et al., 1998]. Another similar scoring was put forward by Naguib et al. who performed a clinical study [Naguib et al., 1999] to identify four risk factors that correlated with the difficult intubation, among which is the modified Mallampati score. As these and many other studies show, the Mallampati



(a) Class 1 (b) Class 2 (c) Class 3 (d) Class 4

Figure 3.1 – Modified Mallampati classification and AAM mask

classification is an essential factor in the difficult intubation prediction, Mallampati score 1 and score 4 showing especially strong correlation with easy and difficult intubation respectively. Therefore, an automatic and objective classification of the modified Mallampati score is an important step in the process of developing an automatic difficult intubation assessment system. This will allow us to eliminate inaccurate classifications or inter-physician variations which are generally due to incorrect points of view.

In this chapter, we propose an effective method to assess the modified Mallampati score of patients from a frontal image of the head of the patient, with the mouth open and the tongue protruding to its maximum. For that purpose, we use active appearance model (AAM) to describe the shape of the opening and the texture of the back of the throat. The most important coefficients of the projection of a new image on the AAM principal components are then used to perform classification using support vector machine (SVM).

The rest of the chapter is organized as follows: section 3.2 describes the proposed methodology, section 3.3 contains information about the dataset and the data collection, section 3.4 details the results we obtain with the proposed algorithm and finally section 3.5 concludes this chapter.

3.2 Methodology

The method proposed in this chapter is based on two main components: we use AAMs to extract features, based on shape and appearance of the buccal cavity, and different SVMs to perform feature selection and classification.

3.2.1 Active appearance models

Active appearance models [Cootes et al., 2001] are statistical models of deformable objects which contain both the shape and texture variation among a set of training images of the object. The training process of AAMs consists first of obtaining statistical shape and texture models separately by applying a principal component analysis (PCA):

$$\boldsymbol{s}(\boldsymbol{\alpha}) = s_0 + \boldsymbol{P}_s\boldsymbol{\alpha} \quad and \quad \boldsymbol{A}(\boldsymbol{\beta}) = \boldsymbol{A}_0 + \boldsymbol{P}_a\boldsymbol{\beta}, \tag{3.1}$$

where $\boldsymbol{s}_0$ and $\boldsymbol{P}_s$ represent the mean shape and the eigenvectors of the covariance matrix of the shape, and $\boldsymbol{A}_0$ and $\boldsymbol{P}_a$ represent those of the texture. In order to obtain a combined model of appearance, the model parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are concatenated and a third PCA is applied to this concatenated vector, as described in equation (3.2).

$$\boldsymbol{s} = \boldsymbol{s}_0 + \boldsymbol{Q}_s\boldsymbol{c} \quad and \quad \boldsymbol{A} = \boldsymbol{A_0} + \boldsymbol{Q}_a\boldsymbol{c}, \tag{3.2}$$

where $\boldsymbol{c}$ is the complete appearance model parameters vector, and $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_a$ are the principal modes of the combined variation, retaining a certain amount of the total variance.

Using this model a new instance of the object can be generated by altering the model parameters $\boldsymbol{c}$. The idea of the AAM search algorithm is then to synthesize a new example by the adjustment of model parameters, and it is generally treated as a minimization problem of the difference between the synthesized image and the original unseen image. We refer the reader to chapter 1, and more specifically to subsection 1.3.1, for a more detailed description of AAM.

In this work, we define an AAM consisting of 12 points located on the lower edge of the upper lip, or the upper incisors, depending on their visibility, and on the line on the back of the tongue, such that the parts defining the modified Mallampati score are included in the object. The region modeled by this AAM is shown by the yellow contour in figure 3.1. The AAM fits perfectly to the Mallampati classification case, not only because it efficiently segments the object and models the shape and texture variations among different subjects, but it also includes certain preprocessing steps such as shape alignment and texture warping which make us invariant to factors like translation, in-plane rotation and scaling.

We have manually annotated 100 images of different subjects and trained an AAM using these manual annotations. Then, we project these manually annotated points and the texture contained inside their convex hull onto the three different eigenspaces defined by the AAM model. We thus obtain for each subject the model parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{c}$, which constitute our complete set of features.

At this stage, we use the manual annotations of the mouth to calculate the model parameters to exclude the effect of model fitting accuracy in the classification process. We thus only use the representation part of the algorithm, and not the search part. In the next chapter of this thesis, using a state-of-the-art landmark detector, i.e. not necessary an AAM, will allow us to automatically segment the contour of the mouth. This will provide full automatization of the system, while keeping the representation used here.

3.2.2 Feature selection and classification

Once we obtain the full set of features (the three different model parameter vectors), we perform a selection of features on these three sets separately. By discarding irrelevant and redundant features, feature selection provides performance improvement in classification. This is due to the fact that the AAM parameters are ordered depending on the ratio of the total variation they explain and since this variation is not necessarily caused by the different Mallampati classes, certain coefficients introduce noise, if taken into account. Feature selection is thus a crucial step in the classification process.

In order to select the most relevant subset of features, we train linear SVMs in a recursive manner, removing one feature at each iteration, in a backward feature elimination manner, similarly to what is done in [Guyon et al., 2002]. Linear SVM is a supervised learning method used for binary classification. The model resulting from a linear SVM is a hyperplane of the

form:

$$\boldsymbol{w} \cdot \boldsymbol{x} - a = 0, \tag{3.3}$$

which maximizes its distance to the nearest training data point of both classes. The normal vector to the hyperplane, $\boldsymbol{w}$, can be seen as feature weights where the highest weight indicates the feature that contributes the most to separating the two classes. At each iteration ordering the features in decreasing order of weight w_i and eliminating the feature with the lowest weight allows obtaining a ranking of the features. Feature selection is then performed by selecting the N first features, as a subset, where $N < p$, the total number of features. As linear SVM is a binary classifier, six different classifiers are trained, in a 1-against-1 fashion, resulting in six different rankings of features.

Then, once we obtain the feature subsets using these rankings, we train six different SVM with radial basis function (RBF) kernel using the publicly available LibSVM implementation [Chang and Lin, 2011]. Once again the SVM are trained in a 1-against-1 fashion as better results are generally obtained by this method, compared to other multi-class strategies such as 1-against-all [Hsu and Lin, 2002]. Details of the cross-validation and parameter optimization are presented in the results section. The final classification of the modified Mallampati score is then obtained by majority voting of these 6 classifiers.

3.3 Dataset

The dataset used is composed of 100 images of different subjects, equally balanced between classes. The images are acquired at the University Hospital in Lausanne (CHUV), and the subjects are actual patients who undergo the regular preoperative assessment for anesthesia prior to their elective surgeries. The recording process of the images is part of a larger project on the automatic assessment of difficult tracheal intubation. The subjects included in this dataset are aged between 24 and 81 and the proportion of female subjects is 39%.

The assessment of the ground truth for the modified Mallampati score is then performed by experienced anesthesiologists only on the basis of these images. The Mallampati classification depends highly on the angle of view of the mouth in the images. The images were taken by trained staff such that the head is positioned to obtain the best visibility of the oropharyngeal features.

3.4 Results and discussion

In this section we report the results of the classification using the leave-one-subject-out cross validation method. For each of the 100 subjects we train six different SVMs, one for each pair of classes. Each time the kernel and regularization parameters of the SVMs are optimized using a 5-fold cross validation on the 99 samples in the training set. The corresponding sample

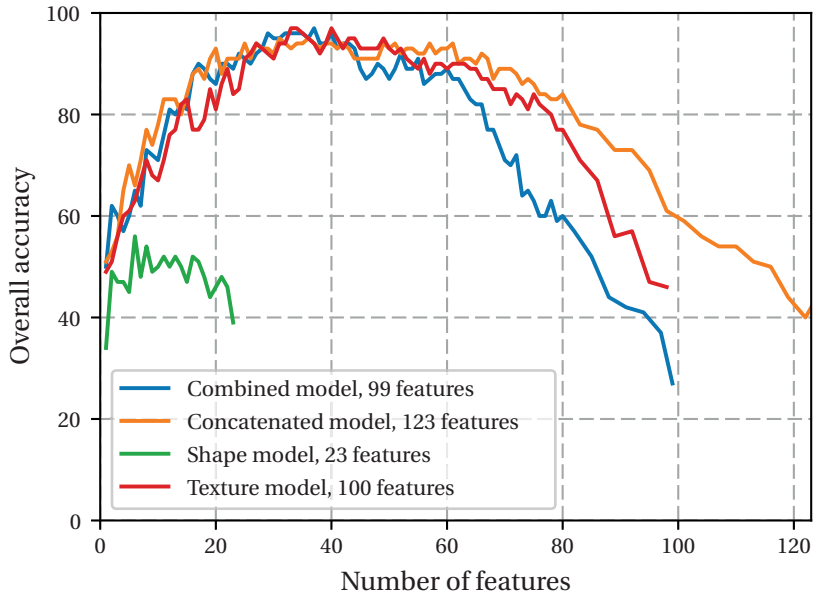


Figure 3.2 – Classification accuracy vs number of features

that was left out is then classified by the six binary SVMs and the final modified Mallampati score is obtained by majority voting.

Feature selection is a key step in the proposed method as explained in subsection 3.2.2. Indeed, we see from the analysis of the feature rankings that, in general, the coefficients corresponding to principle modes explaining a very small portion of the total variance are assigned higher weights. The optimal number of features used by each of the six SVMs is experimentally determined by comparing the overall accuracy obtained by using different numbers of features. Figure 3.2 shows the classification accuracy with respect to to the number of features.

We have performed the tests using the coefficients obtained from the shape model, texture model, and the combined appearance model separately to identify which type of features is the most efficient in discriminating the different Mallampati classes. For each model we keep a number of principal components explaining more than 99.99% of the total variance, resulting in a total of 23 shape features, 100 texture features, and 99 combined features, which are then ranked using the linear SVM method explained in subsection 3.2.2.

Intuitively, the information about the modified Mallampati score is contained mainly in the texture, rather than the shape of the mouth opening. This hypothesis is confirmed by the poor results obtained when using only the coefficients $\boldsymbol{\alpha}$ modeling the variations in the shape. Conversely, using only the coefficients $\boldsymbol{\beta}$ leads to performance of the same quality as using the coefficients $\boldsymbol{c}$, modeling the complete appearance. It can thus be concluded that taking into account the shape does not help to improve the classification performance, as shown in

Table 3.1 – Confusion Table, OA=0.979

	1	2	3	4	
1	21	2	0	0	0.913
2	0	25	0	0	1
3	0	0	24	0	1
4	0	0	0	25	1
	1	0.926	1	1	

figure 3.2.

The best classification performance is obtained using 33 features of the texture model. Table 3.1 presents the confusion table for the corresponding leave-one-out cross validation test. The classification of 3 of the 100 samples was ambiguous due to an equal number of votes in the majority voting scheme. These samples are discarded in the calculation of the final accuracy and not included in the confusion table. In order to avoid such ambiguities, a probabilistic weighting of each classifier in the voting scheme can be used. 95 of the rest of the 97 are correctly classified, corresponding to a 97.94% overall accuracy and 100% recall and precision for Mallampati class 4, which is an important indicator of difficult intubation.

3.5 Conclusion

In this chapter, we proposed an AAM based method to assess the modified Mallampati score of patients from an image of the mouth cavity. We selected the relevant features obtained by the AAM using linear SVM and obtained the classification by majority voting of six different binary SVM classifiers. We performed tests on images of 100 patients and showed that with the optimal number of features we can correctly classify 95% of the total samples, taking into account the 3 samples that were ambiguously classified.

To the best of our knowledge this is the first work proposing an automatic system to assess the modified Mallampati score from images. The modified Mallampati score is often criticized for the lack of objectivity in the way practitioners assess it, especially due to the angle of view. This leads to different scores on the same patient, when examined by different practitioners. In a future work, the proposed image based method can be extended to analyze videos and will allow objectively assessing the modified Mallampati score. In the scope of this thesis, and difficult tracheal intubation prediction, we consider this work as encouraging preliminary results on the subtask of Mallampati classification. This work thus provides an essential element to be integrated into an automatic difficult intubation assessment system, as presented in the next chapter.

4 Automatic prediction of difficult tracheal intubation

4.1 Introduction

In this chapter, we describe a clinical application of facial image analysis to detect morphological traits related to difficult intubation, hypothesizing that advanced facial image analysis methods could improve the prediction of difficult intubation and identify relevant characteristics helping the prediction.

Our proposed method has been developed and validated using more than nine hundred patients. It does not require any medical history or measurement on the patient other than frontal and profile photographs, making it practical even for untrained personnel. The processing of the photographs is completely automatic and does not require any manual initialization. In order to assess its performance in a real-world scenario, we present results including all levels of difficulty and not only very easy and difficult patients. We demonstrate that the proposed method performs as well as state-of-the-art multifactorial tests performed manually by experienced anesthesiologists.

The rest of this chapter is organized as follows: the data collection process and setup is described in section 4.2. In section 4.3, we describe the face models training and fitting processes as well as the learning process. The results obtained are presented in section 4.4 and compared to diagnosis based prediction results. Finally, conclusions and a discussion of future research topics are given in section 4.5.

4.2 Data Collection

Since March 2012, at the University Hospital in Lausanne (CHUV), adult patients, undergoing general anesthesia requiring tracheal intubation and related to any type of elective surgical procedures except obstetric and cardiac surgery, have been preoperativelly recruited. The study has been approved by the Human Research Ethics Committee (Ethical approval number 183/09, Chairperson Prof R. Darioli) from the Ethical Committee of the Canton of Vaud,

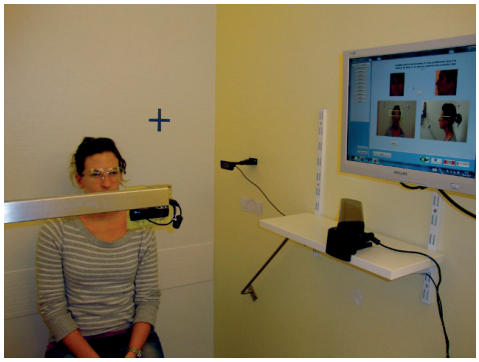


Figure 4.1 – Photo booth at CHUV

Switzerland. Each patient gets appropriate information about the research by the anesthesiologist during the preoperative consultation and gives his or her written consent to participate in the study.

4.2.1 Setup

We developed and set up a *photo booth-like* equipment, depicted in figure 4.1, in the surgical pre-hospitalization center to collect multi-modal data on recruited patients. These data include frontal and profile photos and videos taken with two HD webcams, one in front and one on the left side of the patient at approximately 40 cm. We also record the voice of the patient and capture depth maps with a Microsoft Kinect®.

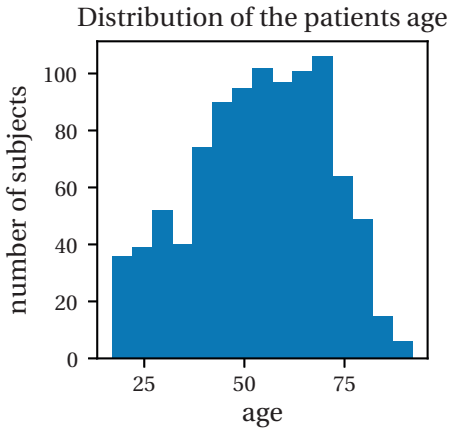
While sitting in the photo booth, the patient is asked to perform different facial motions as well as head motions. Those include: neutral expression, opening the mouth, sticking the tongue out, lateral rotation and vertical extension of the head. A graphical user interface, developed on Matlab, allows an operator to guide the patient through the different poses he has to take and to capture the data at the appropriate moment.

4.2.2 Demographics

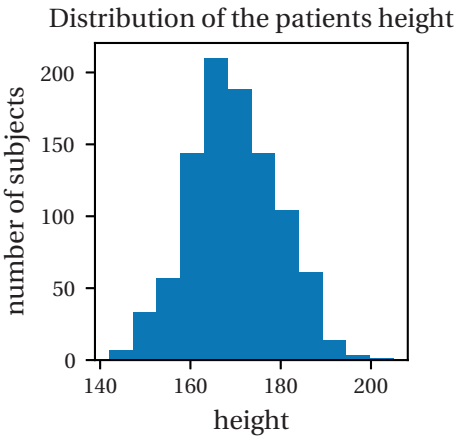
We also collect patient demographics such as age, gender, weight, height and presence of denture during the preoperative anesthesia consultation. Details of peroperative airway management by the in-charge anesthetist are introduced after the operation in a dedicated database. These include information on ease of face-mask ventilation, laryngoscopic grade [Yentis and Lee, 1998] with an appropriate size MacIntosh blade, years of training of intubator, where a minimum of 2 years training in anesthesia was mandatory, lifting force necessary for intubation, either normal or increased, usage of accessory means such as external laryngeal manipulation, intubation bougie, stylet, or video-laryngoscopic equipment, and injuries related to airway management. Number of airway providers and number of intubation trials are also recorded. The intubation difficulty scale (IDS) [Adnet et al., 1997] is routinely calculated.

	Mean [min,max]
Age	53 [17, 92]
Height [cm]	169.5 [142,205]
Weight [Kg]	76.8 [40,160]
Gender [M/F]	488/482
Total	970

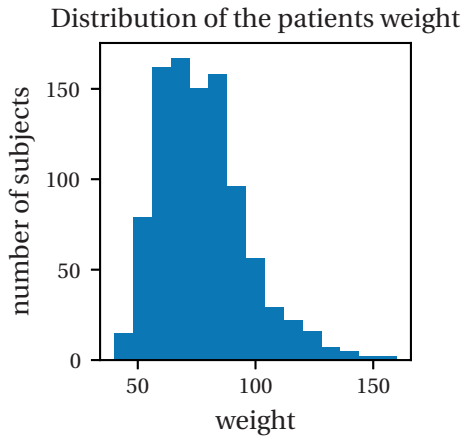
(a) Patients' population metadata



(b)



(c)



(d)

Figure 4.2 – Patients' population metadata and histograms of (b) patients' age (c) patients' height (d) patients' weight

This information allows obtaining a ground truth for the intubation difficulty.

In the two years period from March 2012 to March 2014, we have recorded 2725 patients. The ground truth is available for 970 of those, as detailed in section 4.3.3. Figure 4.2 shows the metadata of the patients' population used in this work.

4.3 Methods

Given a set of images for each patient, we make use of facial image analysis methods in order to extract meaningful features from the face and neck. The location of the face in the image is provided by a face detector and used as initialization for the face alignment algorithm, which provides the localization of the facial landmarks. The features include simple distances between selected facial landmarks as well as information on the global shape or texture variation of the head. In a second step, the statistical relevance of those features is computed in order to discover which of them are relevant in the scope of *prediction of difficult intubation*. The most relevant features are then fed to a classifier. The classifier *learns* how to discriminate between easy, intermediate and difficult to intubate patients.

4.3.1 Detecting the face and tracking the landmarks

Facial image analysis methods often include two main parts: first we need to determine automatically the rough location of a face in the image using a *face detector*, then precise locations of each landmark are found by accurately *fitting a model* of the face on the image. Features are computed using individual landmark positions as well as their global configuration and finally a classifier is trained according to the task. For more details about facial image analysis pipeline, see chapter 1.

Face detector

In order to initialize the fitting of the face model, both the rough location of the face in the image, as well as its scale, need to be determined.

We use Yang and Ramanan's Parts Based Detector [Yang and Ramanan, 2011] in order to detect the face in the images. This method is a general, flexible mixture of parts model able to capture contextual co-occurrence relations between parts, augmenting standard spring models that encode spatial relations. It has been shown to perform very well on face detection [Zhu and Ramanan, 2012] and to be particularly reliable for extreme head poses. The good flexibility of the method allows us to train a single detector for all frontal images, even though the patients are performing very different facial motions, such as opening the mouth widely or sticking out the tongue. An additional detector is trained for profile images as many parts of the frontal images are not visible in the profile images. We use a manually annotated subset of our data to train both detectors. For the frontal detector, the training set consists of 406 annotated images including neutral face, mouth open and tongue out images. Both the original image and the horizontal flip of the image are used. For the profile detector, the training set consists of 134 annotated images.

The frontal face detector performs very well and detects 100% of the frontal faces in the 2910 images of the 970 patients performing all facial motions. This set includes 2553 unseen images,

i.e. not used for training the face detector. The profile face detector, on the other hand, fails to detect the face of only 4 patients, which are removed from the final analysis, reaching a detection rate of 99.56% on unseen images. The detection of the face provided by the face detector is then used to initialize the fitting process of the face model.

Face model for the image alignment problem

Finding the precise location of each pre-defined facial landmark in a new, unseen image is considered as an *image alignment problem*. Image alignment is the process consisting of rigidly moving and non-rigidly deforming a *template* to minimize its *distance* to a query image. Image alignment process is characterized by three elements: *template representation*, *distance metric* and *optimization scheme*.

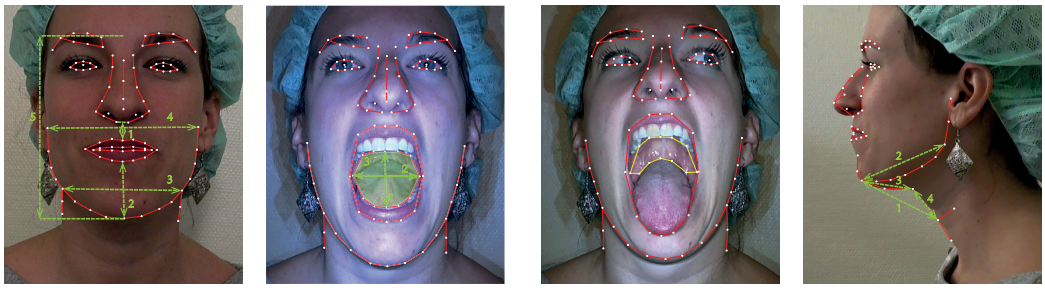
In this work, we follow the image alignment method described in [Xiong and De la Torre, 2013]. The template is non-parametric and consists of scale-invariant feature transform (SIFT) features [Lowe, 2004] extracted from patches around each landmark. This non-parametric shape model is able to better generalize than other parameterized appearance models (PAMs) in unseen situations and this representation is robust against changes in illumination. The squared difference between the SIFT features values computed in the aligned image and in the template is used as the distance metric. This results in the following minimization problem over $\Delta \boldsymbol{s}$:

$$f(\boldsymbol{s_0} + \Delta \boldsymbol{s}) = \| \boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s_0} + \Delta \boldsymbol{x}) - \boldsymbol{\phi}_* \|_2^2 , \tag{4.1}$$

where $\boldsymbol{s_0}$ is the mean shape, $\Delta \boldsymbol{s}$ is the update of the shape, $\boldsymbol{I}$ is the image, $\boldsymbol{\Phi}$ is a non-linear feature extraction function, in our case the SIFT features, and $\boldsymbol{\phi}_* = \boldsymbol{\Phi}(\boldsymbol{I}, \boldsymbol{s}_*)$ represents the SIFT values in the manually labeled landmarks.

The supervised descent method (SDM) optimization scheme, thoroughly described in [Xiong and De la Torre, 2013], learns a series of descent directions and re-scaling factors, equivalent to the Hessian in the case of Newton's method, such that it produces a sequence of updates $\boldsymbol{s}_{t+1} = \boldsymbol{s}_t + \Delta \boldsymbol{s}_t$ starting from $\boldsymbol{s}_0$ that converges to $\boldsymbol{s}_*$ in the training data. $\boldsymbol{s}_0$ is the initial configuration of the landmarks provided by the face detector which corresponds to an average shape, scaled and translated, and $\boldsymbol{s}_*$ is the correct configuration of the landmarks, generally obtained by manual annotations of the images.

Definition of the templates In the scope of this work, we define one template per facial motion, necessary to get accurate landmark positions on photos with different facial motions. In order to train these models, we have defined one neutral and frontal template with 99 points, two different frontal 99 points templates with large facial motions, one with the mouth open and the second with the mouth open and the tongue out, and one profile template consisting of 52 points. We then manually annotated images for each of those templates to train the face



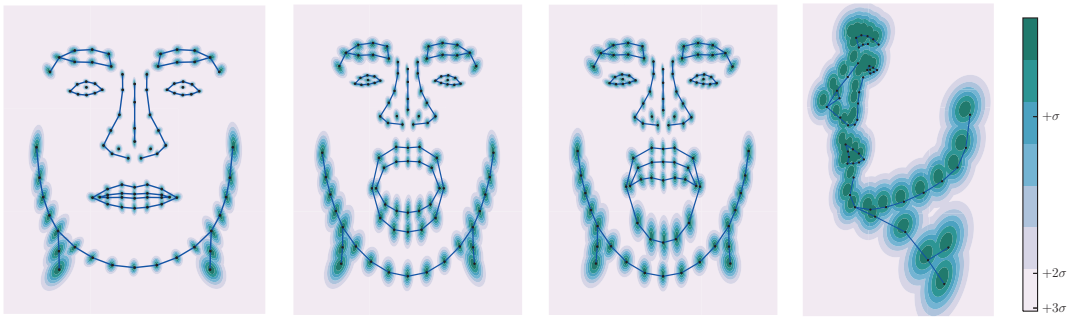
(a) (b) (c) (d)

Figure 4.3 – Details of the four templates, each corresponding to a facial motion: (a) frontal, neutral, 99 points (b) frontal, mouth open, 99 points (c) frontal, tongue out, 99 points (d) profile, neutral, 52 points. In green, the anatomical and morphological features described in section 4.3.2.

model described above. Figure 4.3 shows the facial landmarks configuration corresponding to each template. The facial landmarks are in white and are linked by red segments, for better visualization.

The template corresponding to a neutral position and neutral expression contains landmarks for each eyebrow, eye, the nose, the mouth, and the chin; it has 99 points in total, as shown in figure 4.3a. It also includes points on the neck in order to assess neck characteristics, such as its width. The two templates corresponding to images with extreme facial motions, i.e. mouth open and tongue out, have the same points as the neutral 99 points template as shown in figures 4.3b and 4.3c. The landmarks defining the internal perimeter of mouth opening follow teeth or lips, depending on what is present in the image. The same set of landmarks was used for assessing the tongue out movement with a segmentation of the oral cavity, allowing grading of an automated modified Mallampati classification, as presented in chapter 3. The segmentation of the oral cavity is shown in yellow in figure 4.3c. For profile images, a template of 52 points was defined and is depicted in figure 4.3d. The points on the jaw and the neck allow assessing jaw movement while performing mandibular movement.

Validation of the face model In order to validate the face model, we use K-fold cross-validation. For each model, the images from one fold are kept for testing the model while the images from all other folds are used to train the model. The greater the number of folds, the more training images are used at each run. The obtained model is then fitted on the annotated images in the excluded fold and the obtained landmark positions are compared to the manual annotations. This procedure is repeated for each fold. This way, the model is tested on each available annotated image. Note that the face detector is first run on the images in order to initialize the face model. We thus test the whole pipeline at once. In order to quantify the evolution of the error with respect to the number of training images, we run this K-fold cross-validation scheme for each model with 2, 3, 4, 5 and 10 folds. These correspond to 50%,



(a) (b) (c) (d)

Figure 4.4 – Distribution of the errors, i.e. differences between the landmark positions obtained automatically and the manual annotations, on each landmark for the four templates: (a) frontal, neutral, 99 points, (b) frontal, mouth open, 99 points, (c) frontal, tongue out, 99 points, (d) profile, neutral, 52 points.

66.6%, 75%, 80% and 90% of the annotations used for training. The total number of annotated images is 150 for each of the frontal models and 92 images for the profile model.

Figure 4.4 shows the distributions of the errors for each landmark and each model, when trained and tested using 10 folds cross-validation, which corresponds to using 90% of the annotations for training. During the testing step, the error between each landmark and the corresponding annotation is computed for each test image. We then report these errors on the mean shape of each model and fit a Gaussian function for better visualization.

The quality of the model varies from one model to the other. The profile model is the least accurate, as shown in figure 4.4d, but is also trained on fewer images. Moreover, the annotations might be less consistent from one training image to the other, due to the increased difficulty of annotating the profile face. The points on the chin and the neck, from the profile model, do not correspond to any salient landmarks on the images, therefore increasing the annotation difficulty, as well as decreasing the face tracker ability to precisely locate these landmarks.

Figure 4.5 shows the mean point-to-point error normalized by the distance between the eyes for the three frontal models. Amongst those, the two models with the mouth open and the tongue out exhibits a larger normalized point-to-point error than the neutral one. Again, the points on the chin and the neck are the less accurate, as shown in figure 4.4. It should be noted that the points around the mouth are reasonably accurate and these are also the most interesting for our application. The points around the eyes are the most accurate, thus making them good candidates for normalization. It can be seen that removing the landmarks from the chin and the neck from the mean computation improves the mean point-to-point error by 15% to 25% depending on the model. Indeed, those landmarks are significantly less accurate than the rest of the model, as discussed earlier. In the final application, all available annotated images are used for training. Thus, the actual performance of the models will be better as they

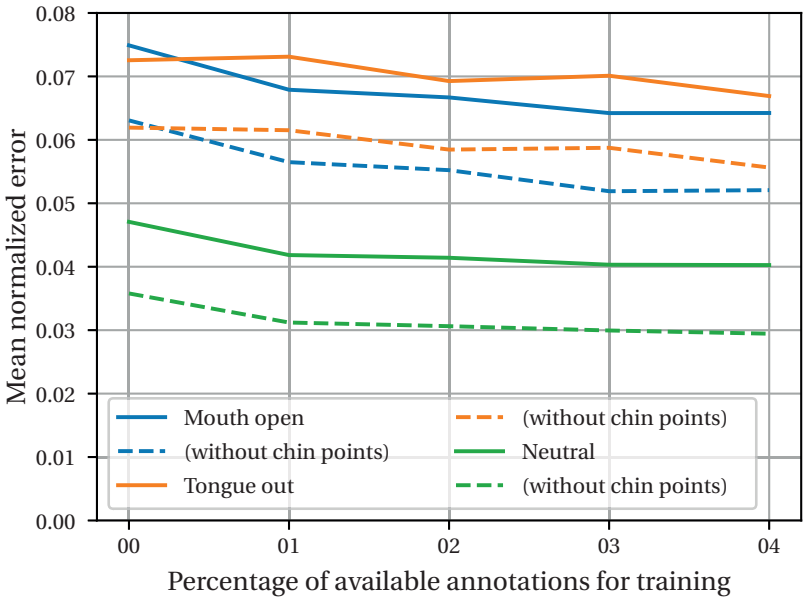


Figure 4.5 – Mean point-to-point error (distance between the landmark positions obtained automatically and the manual annotations) normalized by the distance between the eyes

will have been trained with more annotated images.

4.3.2 Computing the features

Most of the anatomical and morphological features of interest consist of distances between landmarks on the face and the neck. The aligned template gives the positions of these landmarks after fitting the face model on the subject image. Specifically, these distances are: the vertical distance between the upper lip and the nose, the vertical distance between the lower lip and the tip of the chin, the width of the neck, the width of the face, and the height of the face, all five computed on the frontal neutral image, as depicted by lines 1-5, respectively, in figure 4.3a. They are the thyromental distance (TMD) in neutral position, the distance between the angle of the mandible and the tip of the chin, the distance between the hyoid bone and the chin, and the distance between the hyoid bone and the thyroid cartilage, all four computed on the profile neutral image, as depicted by lines 1-4, respectively, in figure 4.3d. Finally, they are the height of the mouth opening, the width of the mouth opening, and the area of the mouth opening, all three computed from the frontal image with the mouth open, as depicted by lines 1-2 and surface 3, respectively, in figure 4.3b. In addition, we compute the distance between the eyes on all frontal images. This distance is used to normalize the features listed above allowing us to be more robust against moderate head pose variations, and to be able to compare them between patients. Indeed, the fact that all patients do not sit at the exact same distance to the camera and do not have the same head pose introduces an important

bias in the features. After normalization, all distances are divided by the distance between the eyes. This one exhibits small variations between subjects, is most likely not correlated with difficult intubation and can be computed reliably from the landmarks around the eyes as they are very accurate.

In addition to the distances between landmarks, we also consider coefficients from a principal component analysis (PCA) on the shape and coefficients from a PCA on the texture, for the inside of the mouth on the frontal model with tongue out, as features. Specifically we compute these coefficients in the following manner:

To compute the PCA-coefficients on the shape, we consider the set of face images used for training, each image having a set of v two dimensional (2D) landmarks, returned by the face tracker, $[x_i, y_i], i = 1, 2, ..., v$. The collection of L landmarks of one image is treated as one observation from the random process defined by the shape model $\boldsymbol{s} = (x_1, y_1, x_2, y_2, ..., x_L, y_L)^T$. Eigenanalysis is applied to the observation set, keeping 98% of energy, and the resultant model represents a shape as

$$\boldsymbol{s}(\boldsymbol{p}) = \boldsymbol{s}_0 + \sum_{i=0}^{n} p_i \boldsymbol{s}_i \ , \tag{4.2}$$

where $\boldsymbol{s}_0$ is the mean shape, $\boldsymbol{s}_i$ is the i^{th} shape basis and $\boldsymbol{p} = (p_1, p_2, ..., p_n)^T$ are the shape parameters.

These parameters $\boldsymbol{p}$ provide information on the global variation of the shape. They are ranked by the value of their corresponding eigenvalue, in a decreasing order, or, similarly by the amount of total variance of the training data that they explain. The first modes of variation explain the bigger amount of total variance and are thus likely to explain the variance of the data due to head pose, gender or other factors that are not significant in the prediction of the difficult intubation. On the other hand, the last ones only explain a small amount of the total variance and merely model the effect of noise in the annotations. Even though not all coefficients are relevant for classification, each of them has the advantage of encoding a variation mode affecting the relative configuration of several landmarks by itself. Thus, by selecting a few, relevant coefficients, we can potentially get information about global configurations of landmarks, or global morphology of the face, correlated with difficult intubation.

To compute the PCA-coefficients on the texture, we first compute a piecewise affine transform between the landmarks segmenting the oral cavity on each image, as shown by the yellow contour in figure 4.3c, and the same landmarks on the mean shape. The texture inside those landmarks is then warped onto the mean shape and normalized to zero mean and unit standard deviation. At training time, the warped and normalized texture from the images in the training set are used to compute a PCA basis. Similarly to the PCA on the shape, the eigenvectors corresponding to the biggest ordered eigenvalues and explaining 75% of the texture variance are kept while the others are discarded. At testing time, the warped and normalized texture from the images in the testing set is then projected on that basis, resulting

in a vector of coefficients used as features. For more details, the reader is referred to chapter 3, in which the same method is used for automatic Mallampati classification.

Section 4.3.3 provides more details about the *feature selection* techniques that have been used to find those relevant coefficients.

4.3.3 Classification

Class definitions

In order to train and test the system, each patient is assigned one of the following labels, considered as ground truth and related to their difficulty of intubation: *easy*, *intermediate* or *difficult*. As no precise definition of the *difficult intubation* has been unanimously accepted, this classification is obtained by combining two complementary definitions, namely the widely accepted definition of the *difficult laryngoscopy*, which considers a laryngoscopy as difficult if the Cormack-Lehane view of the larynx is graded III or IV [Cormack and Lehane, 1984] and the definition based on the *IDS* proposed by Adnet [Adnet et al., 1997], which considers an intubation as difficult if the IDS is greater than 5. We refer the reader to chapter 2, and more specifically 2.1 for complete definitions and a discussion about the Cormack-Lehane classification of the laryngoscopic view and Adnet's IDS. We use this broader definition of the difficult intubation in order to remove, as much as possible, the subjectivity of using only the laryngoscopic grade, while still assigning laryngoscopic grades III and IV to the difficult class. More specifically, the class labels are defined as follows:

easy $IDS = 0$, this implies a laryngoscopic grade of I and a successful intubation at the first attempt;

intermediate $0 < IDS \leq 5$ and laryngoscopic grade smaller than III;

difficult $IDS > 5$, or laryngoscopic grade of III or IV.

Out of the 2725 patients who have been recorded, information allowing to compute the IDS is available for 34.4% and laryngoscopic grade for 51.4%. For the rest of the patients, the anesthesiologist may not have filled the ground truth form that allow us to collect these data. Table 4.1a shows the distribution of patients according to the laryngoscopic view for all recorded patients and for the subset of patients with available ground-truth and face detection. The laryngoscopic view was observed by the anesthesiologist at the intubation time. It should be noted that the classes are largely unbalanced, higher laryngoscopic grades being rarely observed which makes the classification task more challenging. Table 4.1b shows the classification of the recruited patients according to their IDS score. The same remark applies regarding high IDS scores.

Table 4.1c shows the distribution of each class according to the classification described above for the 966 patients used in total. The *easy*, *intermediate* and *difficult* labels are used as ground

Table 4.1 – Distribution of the patients according to different criteria used to define the ground-truth: (a) Patients laryngoscopic grade (LG) distribution as observed by the anesthesiologist at intubation time (b) Patients IDS score distribution (c) Final ground truth labels distribution.

(a)

	recorded patients		966 used patients	
LG		[%]		[%]
1	1083	77.30	708	73.29
2a	208	14.85	158	16.36
2b	57	4.07	47	4.86
3	40	2.85	40	4.14
4	13	0.93	13	1.35

(b)

IDS score	Difficulty		[%]
0	Easy	561	59.87
$0 < IDS \leq 5$	Slight Difficulty	353	37.67
$5 < IDS$	Moderate to Major	23	2.46

(c)

Difficulty		[%]
Easy	561	58.07
Intermediate	345	35.72
Difficult	60	6.21

truth. Note that this does not directly correspond to the IDS because 8 patients with $IDS \leq 5$ have a laryngoscopic grade greater than II and are labelled as *difficult* and 29 other patients with a laryngoscopic grade greater than II have missing IDS score.

Data partition for training and testing and class imbalance problem

The feature selection, the choice of the hyper-parameters, and the training of the classifier are performed on a subset of patients: the *training* set. A distinct subset of patients is then used to test the classifier and compute the different metrics assessing its performance: the *testing* set. The partition of the original data into these two subsets is random but the original distribution of classes is maintained; we perform *stratified* partitioning. In order to compute proper statistics for the results, these training and testing sets are generated several times, each time with different random partitions of the patients.

Note that both the training and the testing set follow the same class distribution as the original dataset. As previously discussed, the occurrence of difficult laryngoscopy has been reported to range from 0.3% to 13% [Naguib et al., 1999]. More recently the occurrence of difficult intubation has been reported between 4.5% and 7.5% in the overall population [Shiga et al., 2005]. In the present dataset, 6.21% of the patients fall in the *difficult* class. From a machine learning point of view, skewed distributions of classes make the learning of concepts more difficult. This is known as the *class imbalance problem*. Even a relatively small imbalance ratio of the order of 10:1, as in our case, is sufficient to hinder the learning process.

Artificially balancing the classes is possible using *sampling methods*. However, those methods present some significant drawbacks [López et al., 2013, He and Garcia, 2009, Galar et al., 2012]. Undersampling from the majority class, or classes, allows reducing the imbalance ratio or even totally compensating for the class imbalance. But removing samples from classes may result in loss of information, thus potentially penalizing the classifier's performance. In the other case, oversampling from the minority classes allow for the same reduction of class imbalance but presents a different drawback. Replicating samples tends to lead to overfitting. Even though more complex techniques exist, several problems prevent from finding a good approximation of the original class density function, for example small disjuncts or class overlapping.

In this work, we consider binary classifiers. To overcome the class imbalance problem, we use the fact that for each sample, *probabilistic classifiers* compute confidence values of belonging to each class. The classifier then usually assigns the most probable label to each sample by maximizing $P(j|x)$, the posterior probability of classifying a sample x as j. Nevertheless, in cost-sensitive learning, given a cost matrix defined as $C(i,j)$ the misclassification cost of classifying an instance from its actual class j into the predicted class i, the minimum expected loss can be determined as:

$$\mathcal{R}(i|x) = \sum_{j \in \{0,1\}} P(j|x) \cdot C(i,j) \,, \tag{4.3}$$

where $\mathcal{R}$ is the Bayes risk and $P(j|x)$ is the posterior probability.

Elkan [Elkan, 2001] showed that modifying the classifier's threshold, in other worlds choosing the positive class if its confidence value is greater than a threshold but not necessarily greater than the confidence value of the other class, has the same effect as sampling in terms of bias but without the drawbacks mentioned above. Thus, defining a threshold θ for the classifier allows compensating for the bias towards the majority class. Specifically, in cost-sensitive learning the optimal threshold θ^* of a classifier with respect to a given cost matrix is defined as:

$$\theta^* = \frac{C(1,0)}{C(1,0) + C(0,1)} \,. \tag{4.4}$$

In binary classification, $C(1,0)$ represents false positive (FP) and $C(0,1)$ represents false negative (FN). The prior probabilities of the negative and positive samples, $p(0)$ and $p(1)$ respectively, are proportional to the number of samples in the original training set. As doubling FN or halving FP has the same effect as doubling $p(1)$, we train the classifier on the complete, unbalanced, training set, and when testing it on the test set, the threshold θ is set to the imbalance ratio between the classes, as described in equation (4.5).

$$\theta = \frac{FP}{FP + FN \cdot \frac{p(0)}{p(1)}} = \frac{1}{1 + \frac{p(0)}{p(1)}} \approx \frac{p(1)}{p(0)} \,, \tag{4.5}$$

where $\frac{p(0)}{p(1)}$ is larger than 1 as the positive class, with the label *difficult*, is the class for which we

have less samples.

As modifying the threshold of the classifier is equivalent to sampling, we compare three methods of choosing this threshold:

- the class imbalance ratio method as described above in equation (4.5),
- minimizing the distance between the corresponding point on the ROC curve and the (0,1) point, i.e. the upper left corner, and
- maximizing the Youden index, i.e. the vertical distance between the corresponding point on the ROC curve and the line of no-discrimination.

The latter two methods use four fold cross-validation on the training set to learn the optimal threshold. In order not to hinder the learning process when training the classifier on an unbalanced set, we use the receiver operating characteristic (ROC) curve and its area under the curve (AUC) as criterion. The ROC curve is generated by plotting the false positive rate (FPR) against the true negative rate (TPR) for all values of the classifier threshold. Independently of what kind of classifier is used, we train it such that the ROC curve generated from the output confidence values maximizes the AUC, since AUC is insensitive to the class imbalance. As a post-processing step, we then compute the threshold to apply on the confidence values in order to obtain the final classification of each sample.

Feature selection and classification

Feature selection is performed on the training set. The goal is to determine which features are the most relevant for difficult intubation prediction. Amongst the complete set of features, only these most relevant features are then used to train the classifier. Reducing the dimensionality of the data, as well as removing noisy, irrelevant features from the data helps improving the classification performance.

Random Forest classifiers provide a feature importance measure which allows for feature ranking and selection [Breiman, 2001]. The feature importance is measured by randomly permuting the feature in the out-of-bag samples and calculating the percent increase in misclassification rate as compared to the out-of-bag rate with all variables intact. From the ranking of the features according to their importance, we only keep the k best and discard all the rest. The parameter k is considered a hyper-parameter and its best value is found using grid-search and K-fold cross-validation on the training set at the same time as the classifier hyper-parameters.

For the final classification, a second Random Forest classifier is used. Random Forest classifiers are known to be less prone to overfitting, due to their use of bagging. Indeed, the training algorithm for Random Forest aims at constructing a forest of trees, where for each tree it randomly samples, with replacement, in the training set and trains the tree, by considering

only a random subset of the features at each splitting node. The hyper-parameters of the classifier are selected using four fold cross-validation on the training set. Specifically, those hyper-parameters are the following: the number of the k best features to keep, in the range 20-180 by step of 10, and the percentage of features to consider at each node when looking for the best split, in the range $0.5\sqrt{N}$ - $2\sqrt{N}$, where N is the total number of features. We use *entropy* as the splitting criterion, as it is less sensitive to class imbalance than the usual accuracy [He and Garcia, 2009]. Our implementation uses Scikit-learn [Pedregosa et al., 2011], a python machine learning library.

4.4 Results

First, we analyze which features are selected and their relevance with respect to existing prediction methods in section 4.4.1. Then, we present two scenarios: an *easy* versus *difficult* classification considering easy control patients and difficult ones in section 4.4.2, as well as a more realistic difficult intubation prediction scenario where all patients are considered in section 4.4.3. The second one would correspond to a real-world scenario where each and every incoming patient gets a prediction.

4.4.1 Analysis of selected features

Figure 4.6 shows histograms of the 5 most selected features by the random forest. These features are the only features selected for all partitions (100 out of 100). The ANOVA F-values and corresponding p-values have been computed for each of those features. Except for the shape coefficient 29 of the model with tongue out, all other features show an F-value > 15 and a corresponding p-value < 10^{-4}. Those are thus informative by themselves, but not the shape coefficient 29, which is informative only in combination with other features. A Gaussian is fitted to the data for each class and each feature for better visualization. Nevertheless, some features do not follow a Gaussian distribution, especially for the *difficult* class.

Except for the height of the mouth opening, all selected features are coefficients from the shape model. The interpretation of these coefficients is not straight-forward as they model global variations of the shape. In order to better understand which morphological characteristics are used to compute the decision, figure 4.7 shows the variations explained by the shape coefficient from the image taken with the mouth open. Figure 4.8 and figure 4.9 show the variations explained by different shape coefficients from the image taken with the tongue out. In figure 4.7 and figure 4.8, the left and right subfigures show the shape corresponding to a value of the coefficient equal to -3σ and $+3\sigma$ respectively, whereas the central subfigure shows the mean shape with the variation of the landmark positions when the coefficient is continuously changed from -3σ to $+3\sigma$.

In figure 4.7, it can be seen that the largest variation in the shape is due to the movements of the landmarks around the mouth. A low value of this coefficient thus represents a configuration

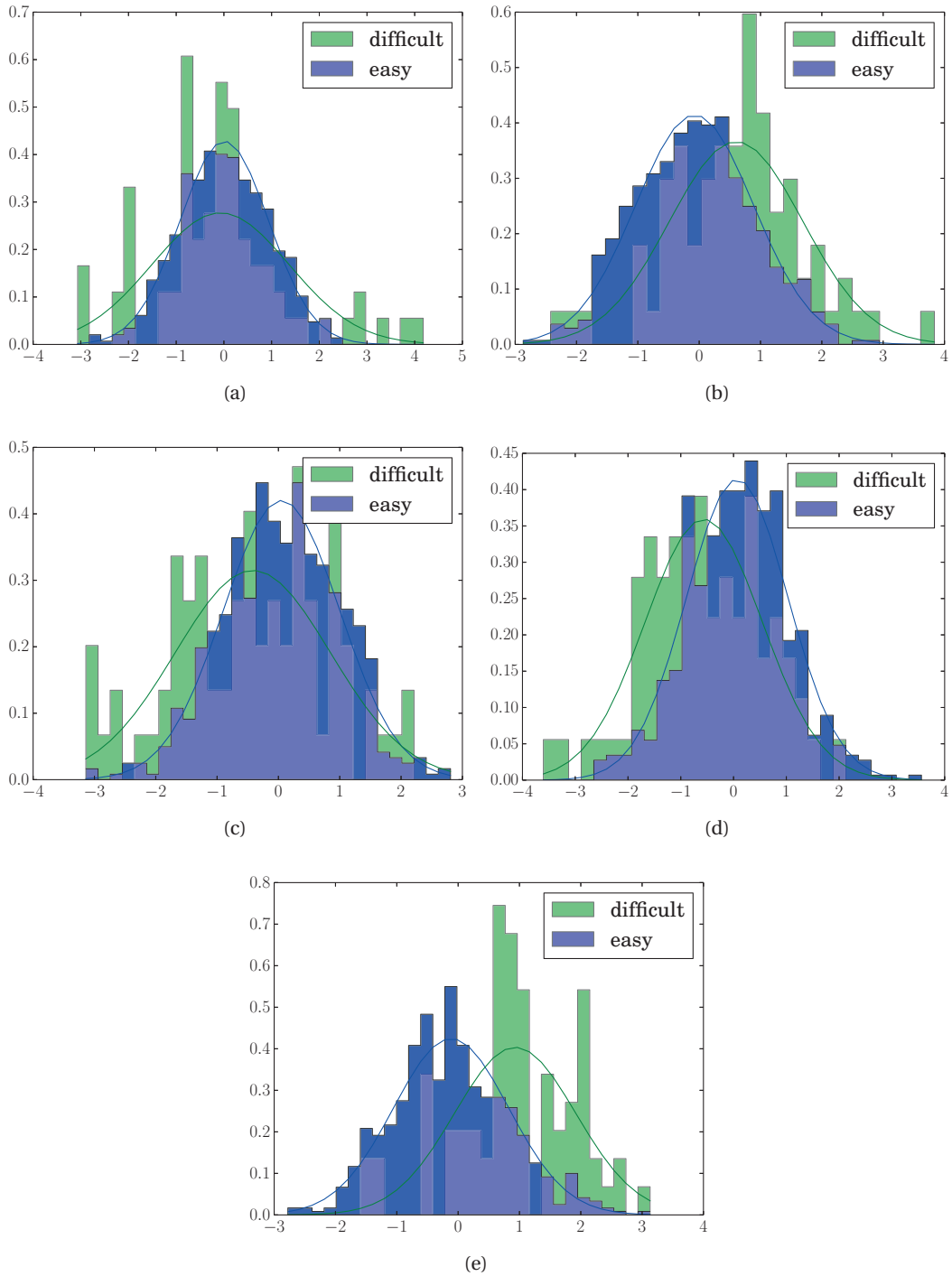
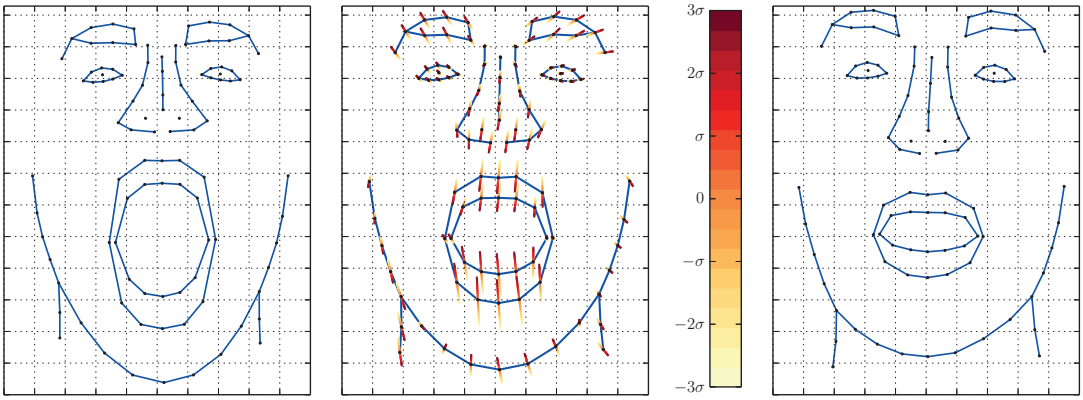
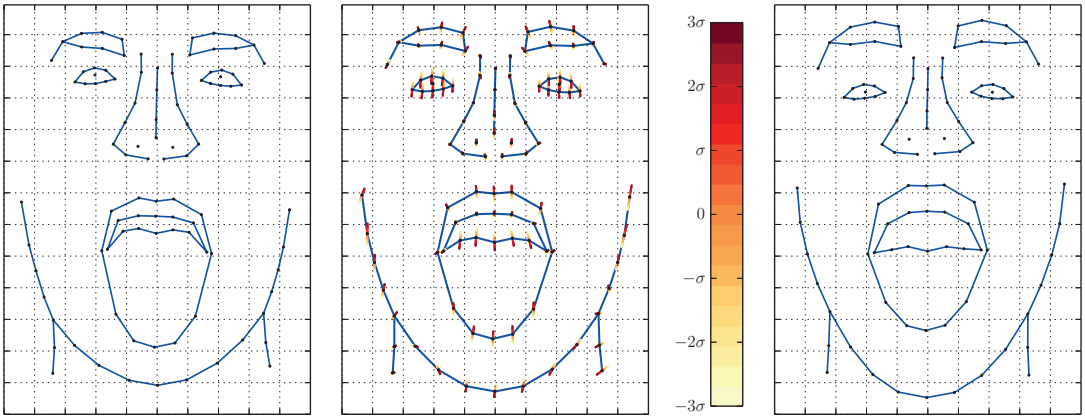


Figure 4.6 – Histograms of the five most selected features: (a) p_{29} (shape coefficient 29) from tongue out image (b) p_2 from mouth open image (c) p_7 from tongue out image (d) height of the mouth opening (e) p_1 from tongue out image



(a) (b) (c)

Figure 4.7 – Mouth open model variations of p_2. (a) -3σ shape (b) variations overlaid on the mean shape (c) $+3\sigma$ shape

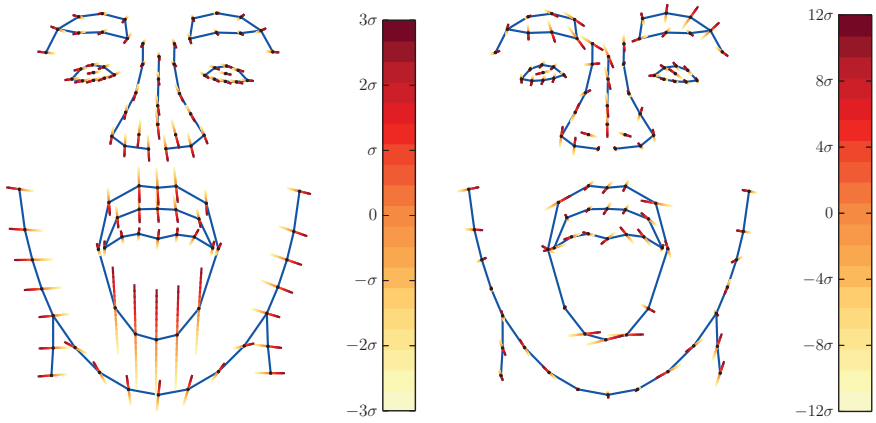


(a) (b) (c)

Figure 4.8 – Tongue out model variations of p_7. (a) -3σ shape (b) variations overlaid on the mean shape (c) $+3\sigma$ shape

of the landmarks corresponding to a face with the mouth widely open, whereas a high value corresponds to a mouth much less open. The classifier thus selected a feature which makes perfect sense as a decreased mouth opening is known as a predictor of difficult intubation by the anesthesiologists, as described in chapter 2 and more specifically in section 2.2.

In figure 4.8 it can be seen that the largest variation in the shape is due to the movements of the landmarks around the eyes, which are not relevant, as well as the movements of the landmarks on the back of the tongue. A low value of this coefficient thus represents a configuration of the landmarks corresponding to a poor visibility of the oro-pharyngeal structures whereas a high value corresponds to a much clearer view of those structures. A small value of that coefficient



(a) (b)

Figure 4.9 – Mouth open and tongue out model variations for (a) p_1 (b) p_{29} (on a different scale)

thus could indicate a bigger tongue, i.e. a tongue with a larger volume. This information is similar to what is indirectly assessed in the modified Mallampati test, as described in section 2.2, and is thus relevant to our classification task, from an anesthesiology point-of-view.

The interpretation of the two coefficients in figure 4.9 is not as straight-forward as for the other coefficients. Those are also less relevant. Indeed, the shape coefficient 1 corresponds to the second largest eigenvalue and thus models a lot of variation in the shape due to many different parameters, whereas the shape coefficient 29 is statistically not relevant by its own with an F-value of 0.43 and a corresponding p-value of 0.51. Those coefficients are merely shown for the completeness of the results.

4.4.2 Easy *vs* difficult classification

In this scenario, we followed the same protocol as Naguib did in his comparative study of four multi-variate difficult tracheal intubation models [Naguib et al., 2006], in which for each *difficult* patient, an *easy* one is selected as *control* patient. In our case, we do not enforce a one to one correspondence, but keep the imbalance between the classes. Removing the *intermediate* patients, we end up with two disjoint classes: the *easy* and the *difficult* patients.

We use 80% of the patients for training and 20% for testing. The partition is repeated 100 times randomly and the results are averaged over those different partitions. This results in 496 training patients (448 *easy* and 48 *difficult*) and 125 test patients (113 *easy* and 12 *difficult*).

The performances of the classifier are reported in table 4.2, along with the results reported in the literature for four manual tests [Naguib et al., 2006], and a previous attempt for semi-automatic difficult intubation prediction from [Connor and Segal, 2011]. We report the mean

Table 4.2 – Comparison of our results on the Easy vs difficult problem with four multivariate tests [Naguib et al., 2006] and a semi-automatic method [Connor and Segal, 2011] in terms of sensitivity (Sens.), specificity (Spec.), and AUC

Model	Sens. [95% CI]	Spec. [95% CI]	AUC
Wilson model[Wilson et al., 1988]	40.2 [30.0, 50.0]	92.8 [88.0, 98.0]	79.0
Arné model [Arné et al., 1998]	54.6 [45.0, 65.0]	94.9 [90.0, 99.0]	87.0
Naguib I model [Naguib et al., 1999]	81.4 [74.0, 89.0]	72.2 [63.0, 81.0]	82.0
Naguib II model [Naguib et al., 2006]	82.5 [73.0, 89.0]	85.6 [77.0, 91.0]	90.0
Connor [Connor and Segal, 2011]	90.0	80.0	84.0
Ours			81.0
class imbalance	79.7 [77.4, 81.9]	67.4 [66.4, 68.4]	
distance to (0,1)	77.1 [74.8, 79.4]	70.6 [69.4, 71.8]	
Youden index	78.9 [76.5, 81.3]	66.7 [64.7, 68.6]	

values of the sensitivity and specificity with their 95% confidence interval (CI).

As can be seen in table 4.2, our fully automatic system achieves comparable performance on the easy vs difficult intubation classification as compared to manual assessment performed by experienced anesthesiologists using state-of-the-art multifactorial tests. In this binary example, the only metric that can be compared directly is the AUC. All other metrics reported can be tuned by varying the threshold of the classifier, depending on the importance given to sensitivity or specificity. This can be seen by comparing the three methods to compute an optimal threshold. The class imbalance method provides the higher sensitivity, which, in this application, is an important metric, as it is critical to detect as many difficult intubations as possible, even at the cost of more false positives.

Figure 4.10 presents the averaged ROC curve over the 100 partitions. In violet, we regenerated the ROC curve corresponding to the validation set in [Connor and Segal, 2011]. We used the values of each samples in the validation set provided in [Connor and Segal, 2011] to compute TPR and FPR for all thresholds. The highlighted performance point on the mean ROC curve has been obtained by setting the threshold of the classifier to the class imbalance ratio. This corresponds to the results reported in table 4.2.

As for comparison with the results reported in [Connor and Segal, 2011], we would like to emphasize that such a comparison would not be a fair one. First, the authors of [Connor and Segal, 2011] trained and tested their system only on male Caucasian patients, in order to limit any potential confounding effects of gender and racial group. We report our results on a much more representative population, as described in figure 4.2. Then, their method is not fully automatic but semi-automatic as it requires manual placement of fiducial markers and manual measurement of the TMD by an anesthesiologist. The number of patients considered to validate their approach is much lower. The authors reported results on a validation set of only 20 difficult and 20 easy patients thus not demonstrating the generalizability of the proposed method. Finally, they state that they perform model selection such that they get the

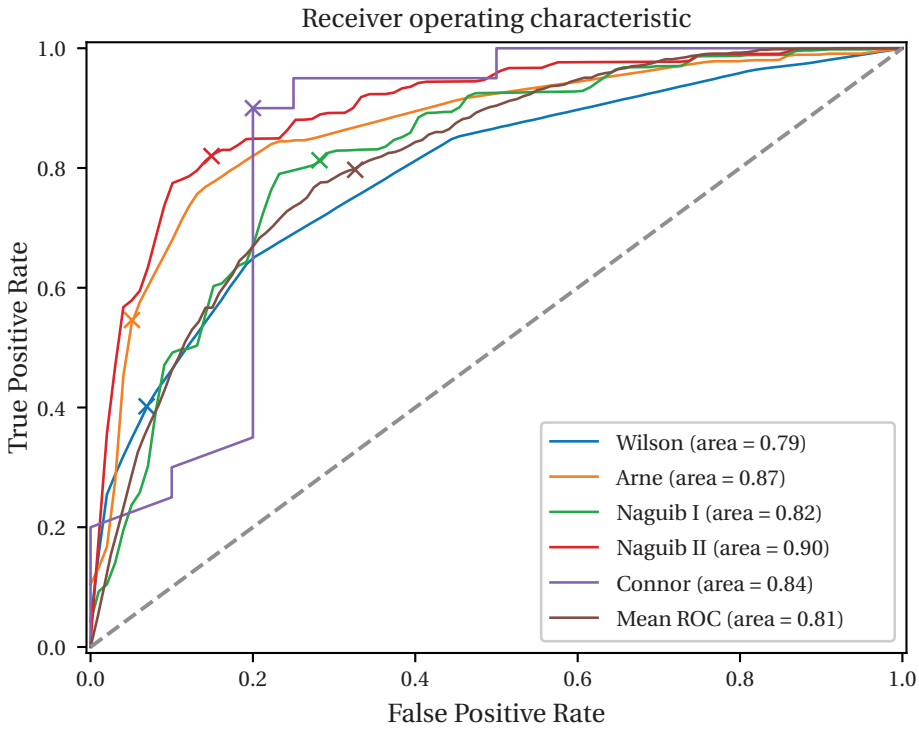


Figure 4.10 – Mean ROC curve for the easy *vs* difficult classification, with performance obtained using the class imbalance threshold method, compared to the ROC curves of four multivariate tests performed manually [Naguib et al., 2006] and the ROC curve obtained on the validation set in [Connor and Segal, 2011]

best product of AUCs on the training and testing sets. Thus, they do not clearly separate the data into training and testing sets and use the testing set to select the model. In addition, they do not perform any kind of cross-validation and demonstrate results on a single partitioning. Methodologically, there is no evidence in their work that similar results would be obtained on an independent test set or a different partitioning of the data. In this work, on the other hand, we present our results on multiple runs, each of them on randomly created independent test sets. Although, in average, our AUC score (0.81) is lower than the AUC calculated on the validation set in [Connor and Segal, 2011] (0.84), our results are better validated in a more generalized way.

4.4.3 Real-world difficult intubation prediction

In the real-world *difficult intubation prediction* problem, the goal is to identify *difficult* to intubate patients from all the others. Considering this task the problem remains a two-class classification problem. Thus, we first group together the *easy* and *intermediate* classes and relabel the new class as *easy*, which *de facto* represents the non-difficult to intubate patients.

Table 4.3 – Comparison of our results on the Real-world problem

Model	Sens. [95% CI]	Spec. [95% CI]	AUC
Real-world			77.9
class imbalance	77.7 [75.7, 79.7]	64.1 [63.2, 65.0]	
distance to (0,1)	72.9 [70.3, 75.5]	68.4 [67.2, 69.5]	
Youden index	74.8 [72.0, 77.5]	65.5 [63.5, 67.4]	

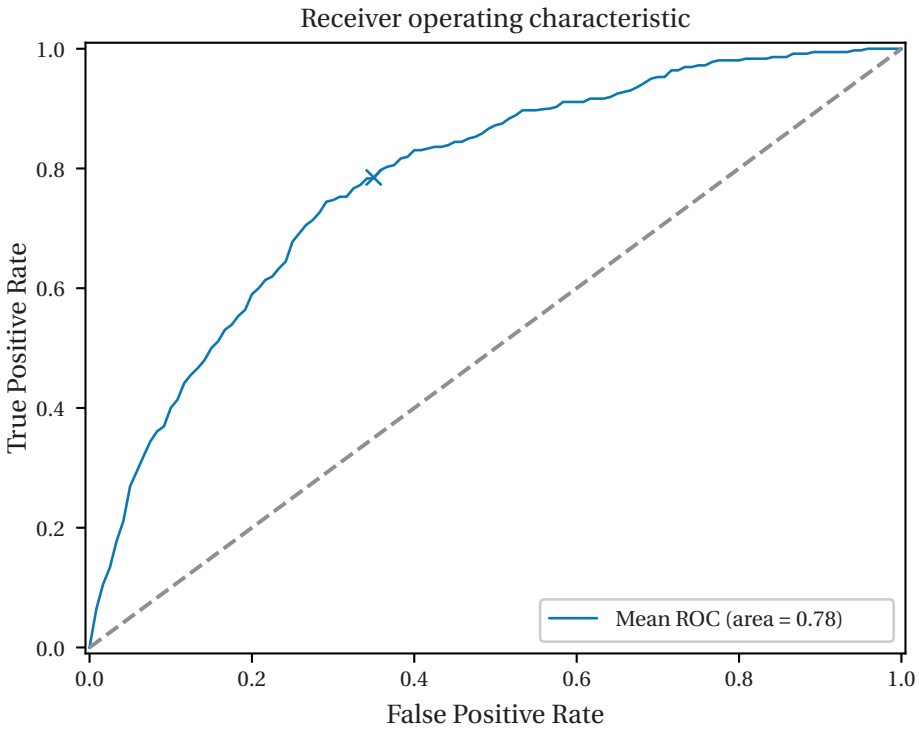


Figure 4.11 – Mean ROC curve for the real-world difficult intubation prediction

When a patient is diagnosed as *difficult*, it sends a strong signal to the anesthesiologists on the potential difficulty of that patient, which is high. Thus, we do not consider only very easy patients as control patients versus difficult ones, but instead we take into account all patients, ranging from very easy to impossible to intubate without gap.

We use 80% of the patients for training and 20% for testing. The partitioning is repeated 100 times randomly and the results are averaged over those different partitions. This results in 772 training patients (724 *easy* and 48 *difficult*) and 194 test patients (182 *easy* and 12 *difficult*). Note that in this case, the class imbalance is more severe, creating an additional challenge to the fact that there is more variation among the samples as compared to the previous scenario. The performances of the classifier are reported in table 4.3. Figure 4.11 presents the averaged ROC curve over the 100 partitions.

As can be seen in table 4.3, the performance of the system drop slightly when considering all patients, without gap between the classes. We observe a 3.1% decrease on the AUC and between -1.2% and -4.2% on the sensitivity and specificity. By considering all patients, the variance of the data is larger. Thus the learning of concepts is hindered as this larger variance can be seen as noise. Moreover, the absence of gap between the classes potentially decreases the class separability, again hindering the learning of concepts. Indeed, the classes become less distinct and when testing on a different dataset than that used for training, the chances are higher that the classes overlap. Note that the definition of the ground truth also has an importance in the performance of the system. More specifically, the subjectivity and poor reproducibility of the Cormack-Lehane grade make the ground truth label less reliable.

4.5 Conclusion

In this chapter, we presented a completely automatic, facial morphometry based method allowing predicting a patient's difficulty of intubation with performance comparable to state-of-the-art medical diagnosis based predictions by experienced doctors. Our method has been validated on more than nine hundred patients, both in a research oriented scenario with only easy and difficult patients and in a real-world oriented scenario where all patients are considered.

The database used in this work is, to the best of our knowledge, the largest database of images, videos, and ground truth data related to endotracheal intubation.

We showed that the learning process takes into account features which have been previously shown to be clinically significant. Of course, the complete decision process takes into account many more variables than the important ones described in this chapter, but it seems reasonable that important clues are also considered.

The open question of how to quantify a difficult intubation remains a penalizing factor for our results. Indeed, the recognized subjectivity, as well as the large variability of the factors taken into account in order to quantify the difficulty of intubation of a patient, create an additional confound for the system. This raises the question of the direct clinical usefulness of such an automatic tool. Yet we demonstrate that it can achieve close to human performance even with such existing limitations. It is thus encouraging to further investigate the usage of facial image analysis in the scope of difficult endotracheal intubation prediction.

A further limitation of the proposed method is its 2D nature. We assume the images to be always frontal and use the distance between the eyes to normalize all morphological features extracted from the face. Clearly, if the head-pose is not perfectly frontal, the normalization is affected and that can potentially introduce noise in the features. With that respect, a three dimensional (3D) model allows to decouple the head-pose and the shape and to remove the effect of head-pose from the features. Moreover, extracting different features independently from different views might be suboptimal. Reconstructing one single 3D shape from multiple

views would also be enabled by a 3D model. In part II of this thesis, we present our work towards such a 3D model.

Due to the rarity of patients difficult to intubate, obtaining a reasonable number of them is a long term procedure. Thus, current and future development include the collection of more data. Another future research axis is to use other modalities that may be indicative of intubation difficulty. For this purpose, we also record the voice of the patient and the depth of the mouth cavity using a Microsoft Kinect®. Further analysis of the data includes the use of these two modalities.

Development of a 3D face model Part II

Overview

The focus of this second part is the development of a new three dimensional (3D) statistical model of the face, which allows to overcome some of the limitations observed in the first part of this thesis. More specifically, a statistical 3D facial model is needed to constrain the inference of the 3D structure of a face from one or several two dimensional (2D) observations. With that respect, it serves as a prior and describe what is a plausible 3D structure of a face. This implies that the data that serve to build the model are critical as the plausibility of the 3D structure of a face will be evaluated based on these.

As we will discuss in this part, essential variations in the 3D structure of a face come from factors linked to the identity, for example the age, gender, and ethnicity, as well as from facial expressions or movements. Thus, sampling from a representative population and including representative facial expressions, depending on the applications, is crucial. Since there existed no existing database of 3D facial scans containing both the same population and the required facial expressions and movements for the application of predicting difficult tracheal intubation described in part I, we recorded our own database, EPFL3DFace. This database will also be useful to the community for applications in 3D facial image analysis.

In order to be able to use the raw 3D facial scans of EPFL3DFace to build a statistical model, these need to be parameterized in a consistent way. Indeed, the vertices of the raw scans do not share any order; the raw scans do not even have the same number of vertices. A consistent parameterization is obtained by nonrigidly registering each scan to a template, thus transferring the parameterization of that template to each scan. Since each scan will then share a common parameterization with the same template, they will also share a common parameterization between them and, thus, be registered.

In this part, we first review available 3D face databases and 3D statistical models of the face, in chapter 5. We also compare our new database to existing ones, in terms of number of subjects, expressions, vertices, and annotated landmarks. This chapter also provides an introduction to important methods and algorithms used in the remaining of this part, namely the Kinect Fusion algorithm and 3D spectral geometry processing methods.

Chapter 6 presents a novel 3D spectral nonrigid registration method using an implicit surface representation and a spectral embedding of the template as deformation model. It also

describes the new database EPFL3DFace in more details and provides results of our proposed spectral nonrigid registration method. Finally, it gives insights on how such a model can be exploited for the problematic presented in part I and similar applications.

The different contributions of this part have been described in a journal article [Cuendet et al., 2017], which has been submitted to IEEE Transactions on Visualization and Computer Graphics and is currently under review.

5 Background

5.1 Introduction

Facial image analysis and synthesis have attracted a significant amount of attention in the last two decades from the computer vision and computer graphics research communities. These two communities have both tackled different but related problems: face recognition [Min et al., 2014, Arar et al., 2012b, Ding et al., 2016], head pose estimation [Fanelli et al., 2012], gaze tracking [Alberto et al., 2012, Arar et al., 2015], visual speech recognition, [Zimmermann et al., 2016] facial expression recognition [Sandbach et al., 2012, Valstar et al., 2015, Jaiswal and Valstar, 2016, Yüce et al., 2013], synthesis of three dimensional (3D) faces [Ichim et al., 2015], facial animation [Weise et al., 2009, Cao et al., 2013, Weise et al., 2011], and face or expression transfer [Thies et al., 2016, Arar et al., 2012a].

The approaches that address these problems can benefit from the availability of low-cost 3D scanners such as the Microsoft Kinect® and take advantage of 3D facial images and 3D face models to avoid limitations inherent to two dimensional (2D) images such as self occlusions or sensitivity to head pose variations. Building a complete 3D face model from the ground up is still very demanding as the amount of data required to obtain a model which takes into account a large amount of variations in terms of identity and facial expressions is high and not easily available from public databases. The variance in appearance is influenced by factors such as age, gender and ethnicity, and when also taking facial expression variations into account, sampling the space of combinations of all these variations simply becomes intractable.

A certain number of databases consisting of 3D representations of the face have been proposed. An important difference between the databases is whether or not the 3D shapes share a common parametrization. Tasks like synthesis of 3D faces or facial animation require a generative model of shapes. These generative models must be learned from a database of consistently parametrized, i.e. registered, instances. Thus, the main challenge in constructing a generative model is to re-parameterize the example surfaces such that semantically corresponding points, e.g. the nose tips or mouth corners, share the same location in the

parametrization domain. Existing 3D face models where 3D scans are registered and statistical analysis is performed include the MPI 3D Morphable Model (3DMM) [Blanz and Vetter, 1999], the multilinear face model [Vlasic et al., 2005], the Basel Face Model, [Paysan et al., 2009], FaceWarehouse [Chen Cao et al., 2014], the Large Scale Facial Model (LSFM) [Booth et al., 2016], the Surrey Face Model (SFM) [Huber et al., 2016] and the Robust Multilinear Model (RMM) [Bolkart and Wuhrer, 2015], but amongst these, only FaceWarehouse and the RMM are trained with a large number of subjects and different facial expressions. These 3D face models are learned from large databases of 3D facial surfaces, containing representative examples spanning the range of variations that the model will be able to capture. As an example, a model learned only from 3D surfaces of neutral faces will not fit well on expressive faces nor be able to capture the variation between a smiling face and a sad face.

In order to allow for statistical modeling, for example with a morphable model, a multilinear model, a blendshape model, etc., the scanned 3D facial surfaces have to be put into dense correspondence by nonrigidly registering the 3D surfaces. The general strategy is for each scan to deform a template, the *floating surface*, or *source* $\mathcal{S}$ such that it matches the scan or *target surface* $\mathcal{T}$. The template parametrization is thus transferred to each of the scans. This nonrigid registration problem is defined by three main elements: a similarity measure, a transformation model, and an objective function. In this thesis, and more specifically in chapter 6, we propose to compute a spectral embedding of the source and use that representation as a transformation model, in order to constrain the possible deformations and enforce smooth deformations.

In this chapter, we aim to provide some background about the different 3D methods that are used in this part of the thesis. We first review existing 3D databases of facial scans in section 5.2. In section 5.3 we give a comprehensive description of the acquisition of 3D scans using a Microsoft Kinect®, and more specifically of the Kinect Fusion algorithm [Newcombe et al., 2011, Izadi et al., 2011], which provides high-quality 3D scans from multiple low-quality depth maps. In section 5.4, we then introduce spectral geometry processing with the aim of providing an intuitive comprehension of the spectral methods on 3D meshes, used in the transformation model of the nonrigid registration method described in chapter 6. We then summarize and conclude this chapter in section 5.5.

5.2 Existing 3D face databases

In the last twenty years, a certain number of databases consisting of 3D scans of the face have been proposed. An important difference between the databases is whether or not the 3D shapes share a common parameterization. Table 5.1 lists databases in which the 3D instances of faces are not registered, *i.e.* do not share a common parameterization. These databases cannot directly be used to build a 3D statistical model, but, with a proper 3D nonrigid registration method, the 3D scans they contain could be registered and subsequently used in a 3D model.

Registered databases, on the other hand, usually serve directly to build a 3D statistical model

Table 5.1 – Comparison of 3D face recognition/verification and head pose databases in terms of number of subjects (subj.), number of expressions (expr.), number of vertices of the aligned surfaces (v.), sensor used for data acquisition, and number of landmarks (landm.). *ND-2006 is a superset of FRGC v.2

Name	subj.	expr.	v.	Acquisition	landm.
XM2VTS	295	Neutral		stereo-based 3D cam	
3DRMA [Beumier and Acheroy, 2001]	100	Neutral		struct.light	
GavabDB	61	4			
FRGC v.2 [Phillips et al., 2005]	(50k rec.)	Neutral, smile		Minolta Vivid 900/910	
BU-3DFE [Yin et al., 2006]	100	Neutral + 6	7.5k	3DMD digitizer	83
ND-2006* [Faltemier et al., 2007]	888	6	112k	Minolta Vivid 900/910	
BU-4DFE [Yin et al., 2008]	101	6	35k	Di3D	83
ETH Face Pose Range Image Data Set [Breitenstein et al., 2008]	26	Neutral	range im.	struct.light [Weise et al., 2007]	
CASIA [Zhong et al., 2007]	123	6		Minolta Vivid 910	
York [Heseltine et al., 2008]	>350	Neutral + 4	5k-6k	3D struct.light cam.	
Bosphorus [Savran et al., 2008]	105	34	35k	Inspeck Mega Capturor II 3D	24
Biwi 3D Audiovisual Corpus of Affective Communication (B3D(AC)2) [Fanelli et al., 2010]	14	15 (speech)		struct.light [Weise et al., 2007]	
Texas 3DFRD [Gupta et al., 2010]	118	Neutral + 3	range im.	MU-2 stereo imaging	25
Photoface [Zafeiriou et al., 2011]	261	>4		4-source PS device	11
UMD-DB [Colombo et al., 2011]	143	Neutral + 3		Minolta Vivid 900	7
UHDB11 [Toderici et al., 2014]	23	Neutral		3DMD	9
Biwi Kinect Head Pose Database [Fanelli et al., 2012]	20	Neutral		Kinect	
BP4D-Spontaneous [Zhang et al., 2013, Zhang et al., 2014]	41	8 spontaneous	30k-50k	Di3D	83
KinectFaceDB [Min et al., 2014]	52	3		Kinect	6

of the face. Blanz and Vetter first introduced the term *morphable model* [Blanz and Vetter, 1999] to describe their parametric face modeling technique based on a large number of 3D face scans. In order to establish correspondence between all individual face scans, they use cylindrical coordinates both for color and geometry information and adapted the optical flow algorithm to compute a vector field of displacement between points [Vetter and Blanz, 1998]. Their method is well suited for data acquired with a 3D scanner using cylindrical coordinates or that can easily be converted to that particular planar representation.

In [Allen et al., 2003], the authors present a template-based nonrigid registration method to compute dense point-to-point correspondence between surfaces with the same overall structure, but substantial variation in shape, such as human bodies. They formulate this as an optimization problem over a set of *per vertex* affine transformations. The objective function includes three terms: a data term defined as the sum of squared distances between spatially close vertices on the source and the target surfaces, a smoothness term which enforces that neighboring affine transformations are as similar as possible and a marker term defined as the sum of squared distances between a set of marker's locations on the template surface and on the target surface. By ensuring the smoothness of the transformations over the surface, they define an *as-rigid-as-possible per vertex affine transform* further constrained with a set of 3D marker locations. By using domain knowledge inherent in the template surface, this method is robust to incomplete surface data and is able to fill in holes or poorly captured parts of the surface.

Vlasic et al. [Vlasic et al., 2005] applied this template-fitting procedure to 3D face scans and described multilinear face models for expression transfer. In [Mpiperis et al., 2008] Mpiperis et al. follow a method similar to [Allen et al., 2003] but add an error term looking for correspondences directed from the target surface to the source and not only in the other direction. They claim that this is important at the beginning of the optimization process when the source is far from the target and it helps avoiding local minima by making the resulting vector field smoother.

Extending the idea of iterative closest point (ICP) [Besl and McKay, 1992] to nonrigid registration and in particular defining optimal steps using a series of stiffness weights to regularize the deformation described in [Allen et al., 2003], Amberg et al. defined the optimal step nonrigid iterative closest point (NICP) [Amberg et al., 2007]. They express the cost function as a least squares problem, thus being able to determine in each step of the algorithm the optimal deformation, in the sense that it exactly minimizes the cost function for fixed stiffness and correspondences.

Further extending the method, Cheng et al. proposed to incorporate a statistical shape prior [Shiyang Cheng et al., 2015] into the fitting procedure of NICP in order to avoid noisy fitting results and even non-face like fitting due to its weak constraint on the shape geometry. The statistical shape prior is a deformable 3D face model [Passalis et al., 2005, Kakadiaris et al., 2007], whose optimal controlling parameters are solved in an alternating manner. Along the

same line, Brunton et al. [Brunton et al., 2014] proposed a detailed review of statistical shape models. They emphasize that to incorporate a statistical shape model to fit to data, instead of a template-based approach with a NICP approach and regularization constraints, can significantly reduce the search space. This results in the ability to reconstruct the underlying shape in the presence of severe noise or occlusions.

Weise et al. [Weise et al., 2009] also followed a NICP approach, optimizing a cost function composed of three terms. Nevertheless, they introduced a combination of point-to-point distance and point-to-plane distance as discussed in [Mitra et al., 2004] in the data-term and expressed the smoothness term as a membrane energy on the displacement vectors, using the standard cotangent discretization of the Laplace-Beltrami operator.

Sumner et al. [Sumner et al., 2007] introduced an embedded deformation model composed of a collection of affine transformations organized in a graph structure. One transformation is associated with each node of a graph embedded in $\mathbb{R}^3$, so that the graph provides spatial organization to the deformations. Each affine transformation induces a localized deformation on the nearby space. That approach was later adapted by Li et al. [Li et al., 2009] to handle motion in the data. This nonrigid registration approach is successfully used for real-time performance-based facial animation in [Weise et al., 2011].

In [Zell and Botsch, 2013] Zell et al. extended the NICP approach to surfaces which cannot be considered near-isometric and for which the closest point correspondences might be invalid by first mapping the source and target surfaces into a simpler space and computing correspondences there. The simpler space is a smoothed, feature-less version of the input models computed by a joint fairing technique based on Laplacian smoothing. To compute correspondences, they iteratively minimize a cost function, which includes three terms: a data term and a marker term, similarly to previously described approaches, and a smoothness term defined as the norm of the Laplacians of vertex displacements, similar to the one used in [Weise et al., 2009].

Recently, Huber et al. released the *Surrey Face Model* (SFM) [Huber et al., 2016], a multi-resolution 3D morphable face model trained with 169 subjects with a neutral facial expression. Their nonrigid registration method was previously described in [Tena et al., 2006] and is an iterative coarse to fine method based on [Zhili Mao et al., 2004]. This method comprises three steps: first landmarks on the source and the target surfaces are brought into correspondence using thin plate spline (TPS) interpolation technique. Then, corresponding points on the source and the target are computed. The search for corresponding closest points takes into account not only the distance between points on the source and the target surfaces but also the angle between their normals, and the difference between curvature shape indices. Finally the positions of the source points are optimized in an *as-rigid-as-possible* fashion.

Bolkart et al. [Bolkart and Wuhrer, 2015] emphasize the chicken-and-egg nature of the problem of training a new statistical face model: given a set of shapes and dense correspondences, a statistical model can be learned and given a representative model, better correspondences

can be computed among a set of shapes. They propose a fully automatic approach to optimize the correspondences for 3D face databases based on multilinear statistical models using groupwise multilinear correspondences [Bolkart and Wuhrer, 2015]. This method measures the model quality and optimizes the registration in such a way that the quality of both the model and the registration improve but an initial registration remains necessary. In their work, they first use a blendshape model to address the expression fitting problem. The 3D blendshapes were manually generated using a commercial software. To further nonrigidly deform the template corresponding to the correct expression, they use an embedded deformation framework [Sumner et al., 2007]. This method was applied to two existing databases of 3D facial surfaces, the Bosphorus database [Savran et al., 2008] and the BU-3DFE database [Yin et al., 2006] and resulted in the Robust Multilinear Model (RMM) [Bolkart and Wuhrer, 2015].

As an alternative to NICP, some methods compute correspondences between two surfaces by embedding the intrinsic geometry of one surface into the other using generalized multidimensional scaling (GMDS) [Bronstein et al., 2006]. The good performance of this kind of methods has been demonstrated for face recognition and are an alternative to deal with variations due to facial expressions [Bronstein et al., 2007b, Bronstein et al., 2007a]. As GMDS methods do not impose that close-by points on one surface map to close-by points on the other, the results are often spatially inconsistent.

In existing 3D facial expression databases, only FaceWarehouse, a 3D facial expression database for visual computing, released by Cao et al. [Chen Cao et al., 2014], has both a large number of subjects and a variety of facial expressions. It consists of registered 3D surfaces of the head of 150 subjects performing 19 facial expressions plus a neutral face. The facial surfaces of the subjects were acquired with a Microsoft Kinect®. To register the 3D scans together, they used a two-step process, close to the NICP methods described above. In the first step, Blanz and Vetter's morphable model [Paysan et al., 2009] is automatically fitted and used as a parametric template. The nonrigid alignment between the fitted model and each of the neutral scans is then refined by allowing the obtained mesh to deform using a Laplacian-based mesh deformation algorithm [Huang et al., 2006]. Finally, the scans containing facial expressions are aligned using a deformation transfer algorithm [Sumner and Popović, 2004] and refined with the same Laplacian-based mesh deformation algorithm.

Table 5.2 provides a comparison of EPFL3DFace, our new face expressions database, with respect to existing 3D face models and databases in which the facial surfaces have been registered and are in dense correspondence with each other. In existing 3D facial expression databases, only FaceWarehouse [Chen Cao et al., 2014] has both a large number of subjects and a variety of facial expressions. In comparison to that database, EPFL3DFace provides additional visemes suitable for visual speech recognition applications, additional facial expressions, and an extreme facial movement. In total, EPFL3DFace contains 35 scans for each subject, whereas FaceWarehouse contains 20 scans. In addition, FaceWarehouse and EPFL3DFace contain subjects from different populations, mostly Asian in FaceWarehouse and mostly Caucasian in EPFL3DFace, and can be considered as complementary in that respect.

Table 5.2 – Comparison of registered 3D face databases and 3D face models in terms of number of subjects (subj.), number of expressions (expr.), number of vertices of the aligned surfaces (v.), sensor used for data acquisition, and number of landmarks (landm.). Note that in the RMM, no new 3D data are recorded, but 3D data from the Bosphorus [Savran et al., 2008] and BU-3DFE [Yin et al., 2006] databases are registered using [Sumner et al., 2007].

Name	subj.	expr.	v.	Acquisition/Source	landm.
3D Morphable Model (3DMM), MPI Tübingen [Blanz and Vetter, 1999]	200	Neutral	≈70k	Cyberware	
Spacetime Faces [Zhang et al., 2004]	1	384	23.728k	Custom structured light scanner	
Multilinear face model [Vlasic et al., 2005]	15 + 16	10 + 10	≈30k	3dMD/3Q's	21
Human Face [Bronstein et al., 2007a]	1	15	≈2k	Custom structured light scanner	
Basel Face Model [Paysan et al., 2009]	200	Neutral	53.49k	ABW-3D	
FaceWarehouse [Chen Cao et al., 2014]	150	20 (47)	11.51k	Microsoft Kinect ®	74
Large Scale Facial Model (LSFM) [Booth et al., 2016]	9663	Neutral	53.215k	3dMD	
Surrey Face Model (SFM) [Huber et al., 2016]	169	Neutral	29.587k[1]	3dMDface	46
Robust Multilinear Model (RMM) [Bolkart and Wuhrer, 2015]	205	23	5.996k	Bosphorus & BU-3DFE	
EPFL3DFace	120	35	11.51k	Microsoft Kinect ®	74

[1] Multiresolution model with different levels of detail and number of vertices: 29.587k / 16.759k / 3.448k

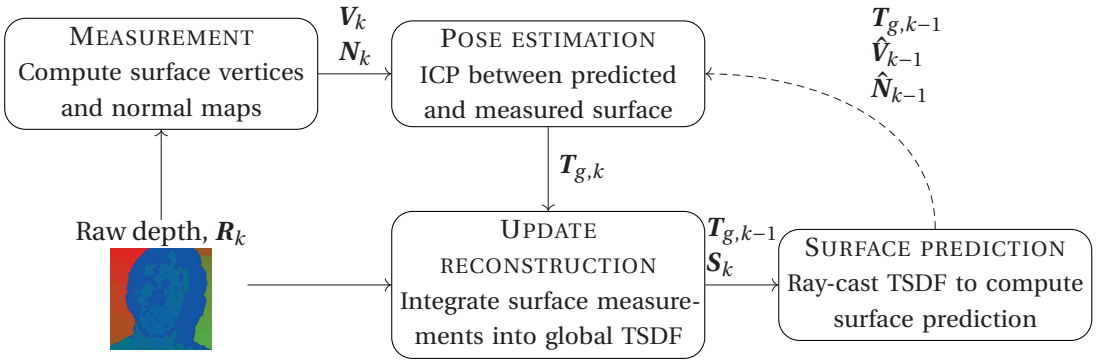


Figure 5.1 – Kinect fusion algorithm scheme

Chapter 6 provides more details about EPFL3DFace.

5.3 Acquisition of 3D scans with the Kinect

Microsoft Kinect® is a low-cost sensor platform that incorporates a structured light based depth sensor. It can generate a 11-bit 640x480 depth map at 30Hz, using an on-board ASIC. Nevertheless, these raw depth maps are very noisy and contain holes where no structured-light depth reading was possible.

Microsoft Research presented the Kinect Fusion algorithm [Newcombe et al., 2011, Izadi et al., 2011], which takes the real-time stream of noisy depth maps from the Kinect and performs real-time dense simultaneous localization and mapping (SLAM). This allows to obtain a consistent 3D scene model incrementally, effectively integrating and denoising the noisy depth maps in a global 3D reconstruction. The system is composed of four main blocs, as described in figure 5.1. From a raw depth map $\boldsymbol{R}_k$, the measurement step computes the vertices' positions $\boldsymbol{V}_k$ and normals $\boldsymbol{N}_k$. These are used to estimate the current pose $\boldsymbol{T}_{g,k}$ of the sensor with respect to the global scene, using an ICP algorithm. The pose of the sensor allows to integrate the raw depth map into the global reconstruction $\boldsymbol{S}_k$, stored as an implicit surface defined in a given volume. Finally, from this implicit global reconstruction, the algorithm predicts a surface by ray-casting the volume containing the implicit surface. Kinect Fusion algorithm is available in an open-source lightweight implementation[2]. We also contributed to that implementation by porting the color integration[3]. In the remaining of this section, we will detail each of these four components.

[2] https://github.com/Nerei/kinfu_remake
[3] https://github.com/gcuendet/kinfu_remake

5.3.1 Measurements

The raw depth map $\boldsymbol{R}_k$, which provides calibrated depth measurements $R_k(\boldsymbol{u}) \in \mathbb{R}$ at each image pixel $\boldsymbol{u} = (u, v)^T$ in the image domain $\boldsymbol{u} \in \mathscr{U} \subset \mathbb{R}^2$, is first filtered with a bilateral filter [Tomasi and Manduchi, 1998]. From the filtered depth map $\boldsymbol{D}_k$, each depth measurement is back-projected to 3D space in order to compute vertices' positions $\boldsymbol{V}_k$ as a point cloud, as described in equation (5.1).

$$\boldsymbol{V}_k(\boldsymbol{u}) = D_k(\boldsymbol{u})\boldsymbol{K}^{-1}\dot{\boldsymbol{u}}\,, \tag{5.1}$$

where $\boldsymbol{K}$ is the camera calibration matrix, which transforms points on the sensor plane into image pixels, and $\dot{\boldsymbol{u}} := (\boldsymbol{u}^T|1)^T$ denotes the vector $\boldsymbol{u}$ in homogeneous coordinates.

From that point cloud, a normal map $\boldsymbol{N}_k$ is computed. It associates a normal vector $\boldsymbol{N}_k(\boldsymbol{u})$ with each depth measurement by computing a cross product between neighbouring vertices, as described in equation (5.2).

$$\boldsymbol{N}_k(\boldsymbol{u}) = \nu\left[(\boldsymbol{V}_k(u+1, v) - \boldsymbol{V}_k(u, v)) \times (\boldsymbol{V}_k(u, v+1) - \boldsymbol{V}_k(u, v))\right]\,, \tag{5.2}$$

where $\nu[\boldsymbol{x}] = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$. The vertex map and normal map are computed in a multi-scale fashion, halving the resolution for each successive level of the pyramid by averaging and sub-sampling the filtered depth map.

5.3.2 Pose estimation

In order to correctly integrate multiple views of the scene into the global 3D reconstruction, it is necessary to compute the current pose of the sensor with respect to the global scene,

$$\boldsymbol{T}_{g,k} = \begin{bmatrix} \boldsymbol{R}_{g,k} & \boldsymbol{t}_{g,k} \\ \boldsymbol{0}^T & 1 \end{bmatrix}\,, \tag{5.3}$$

where $\boldsymbol{R}_{g,k}$ is the rotation component of the pose and $\boldsymbol{t}_{g,k}$ is its translation component.

An ICP algorithm is used to estimate the sensor's pose at each frame with respect to the current global reconstruction. A fast projective data association algorithm [Blais and Levine, 1995] is used to obtain correspondences and the pose is optimized with respect to the point-to-plane error metric, as described in equation (5.4).

$$\boldsymbol{E}(\boldsymbol{T}_{g,k}) = \sum_{\boldsymbol{u}\in\mathscr{U}} \left\| \left(\boldsymbol{T}_{g,k}\dot{\boldsymbol{V}}_k(\boldsymbol{u}) - \hat{\boldsymbol{V}}^g_{k-1}(\hat{\boldsymbol{u}})\right)^T \hat{\boldsymbol{N}}^g_{k-1}(\hat{\boldsymbol{u}}) \right\|_2 . \tag{5.4}$$

Key points of the method are the fact that all vertices of the depth map are used to compute the pose, and not only a limited subset, as is generally the case, and the fact that the pose of the sensor in the current frame is computed with respect to the global reconstruction available so far, and not the previous frame. These are made possible by an efficient implementation

on the graphics processing unit (GPU) and the high frame-rate of the algorithm, limiting the motion from one frame to the other.

5.3.3 Reconstruction update

Each consecutive raw depth map $\boldsymbol{R}_k$, with its associated sensor pose estimate, $\boldsymbol{T}_{g,k}$ is integrated incrementally into one single 3D reconstruction $\boldsymbol{S}_k$ using a discrete volumetric truncated signed distance function (TSDF) [Curless and Levoy, 1996]. We denote the global TSDF that contains a fusion of the registered depth measurements from frames $1,\ldots,k$ as $\boldsymbol{S}_k(\boldsymbol{p})$ where $\boldsymbol{p} \in \mathbb{R}^3$ is a global frame point in the 3D volume to be reconstructed.

In this discrete volume, each voxel stores a running weighted average of its distance to the assumed position of a physical surface. This can be seen as de-noising the global TSDF from multiple noisy TSDF measurements. More specifically, two components are stored in each voxel of the TSDF: the current truncated signed distance value $F_k(\boldsymbol{p})$ and a weight $W_k(\boldsymbol{p})$. The expression of the truncated signed distance value $F_k(\boldsymbol{p})$ is given in equation (5.5).

$$
\begin{aligned}
F_{R_k}(\boldsymbol{p}) &= \Psi\left(\lambda^{-1}\|\boldsymbol{t}_{g,k} - \boldsymbol{p}\|_2 - R_k(\boldsymbol{x})\right), \\
\lambda &= \left\|\boldsymbol{K}^{-1}\dot{\boldsymbol{x}}\right\|_2, \\
\boldsymbol{x} &= \left\lfloor \pi\left(\boldsymbol{K}\boldsymbol{T}_{g,k}^{-1}\boldsymbol{p}\right)\right\rfloor, \\
\Psi(\eta) &= \begin{cases} \min(1, \frac{\eta}{\mu})\mathrm{sgn}(\eta) & \text{iff} \quad \eta \geq -\mu \\ null & otherwise \end{cases},
\end{aligned}
\tag{5.5}
$$

where $\boldsymbol{q} = \pi(\boldsymbol{p})$ performs perspective projection of $\boldsymbol{p} \in \mathbb{R}^3 = (x, y, z)^T$ including dehomogenisation to obtain $\boldsymbol{q} \in \mathbb{R}^2 = (\frac{x}{z}, \frac{y}{z})^T$, $\lfloor \cdot \rfloor$ is a nearest neighbor lookup, and, thus, $\boldsymbol{x}$ is the nearest pixel coordinate of where $\boldsymbol{p}$ would be projected on the image. $\Psi(\eta)$ is the truncation function of the TSDF, truncating $|\eta| > \mu$, where μ is an estimate of the uncertainty on the depth measurement. The expression of the weight $W_k(\boldsymbol{p})$ is given in equation (5.6).

$$
W_k(\boldsymbol{p}) = \frac{\cos(\theta)}{R_k(\boldsymbol{x})}, \tag{5.6}
$$

where θ is the angle between the associated pixel ray direction and the surface normal measurement.

5.3.4 Surface prediction

At this point, the algorithm is integrating successive depth maps into a common volumetric implicit representation of the 3D scene. It is thus possible, at each step k, to compute a dense surface prediction by rendering the surface encoded in the zero level-set $F_k = 0$ of the TSDF. This dense surface is stored as a vertex map $\hat{\boldsymbol{V}}_k$ and a normal map $\hat{\boldsymbol{N}}_k$ and is used in the next

sensor pose estimation step, as illustrated in figure 5.1. In order to render the surface, a per pixel raycast is performed [Parker et al., 1998].

5.4 Spectral geometry processing

In this section, we aim to give an intuitive comprehension of what is spectral geometry processing, and more specifically spectral mesh processing. Since the seminal paper of Taubin [Taubin, 1995], where spectral analysis of mesh geometry is used to describe mesh smoothing as a low-pass filtering operation, this framework has been used in many different applications and we do not aim to provide a complete review of all of them. We refer the reader to the excellent introductory SIGGRAPH course of Lévy and Zhang [Lévy and Zhang, 2010] or Botsch et al. book *Polygon Mesh Processing* [Botsch et al., 2010]. For more in depth coverage of the topic, see the survey of Zhang et al. [Zhang et al., 2010] or the survey of Botsch and Sorkine [Botsch and Sorkine, 2008].

In a nutshell, the framework of spectral methods for mesh processing is the following: the eigendecomposition of a matrix representing a discrete linear operator, based on the topological or geometric structure of the mesh, is performed and the resulting eigenvalues and eigenvectors are used as a new representation of the underlying mesh. The reason why these methods are referred to a spectral might not be evident and comes from the relationship between this type of spectral processing and the Fourier transform. We introduce spectral mesh processing and its link to the Fourier transform in subsection 5.4.1. Then we provide an overview of the challenges in discretizing the continuous Laplace operator in subsection 5.4.2 and present an efficient method to compute the eigenstructures of large matrices in subsection 5.4.3.

5.4.1 Link with the Fourier transform

The mesh vertex coordinates $\boldsymbol{v}_i = (x_i, y_i, z_i)^T$ can be considered as a 3D signal defined over the underlying mesh graph. That is how Taubin [Taubin, 1995] first introduced the use of mesh Laplacian operators in his seminal paper. The classical Fourier transform of a periodic one dimensional (1D) signal can be seen as the decomposition of that signal into a linear combination of the eigenvectors of the 1D Laplacian operator. A proof of that claim can be found in Jain's classic text on image processing [Jain, 1989]. Similarly, defining a discrete Laplace operator on the mesh and projecting the mesh vertex coordinate signal onto the eigenvectors of that Laplacian allows to extend the notion of Fourier transform to the manifold setting.

The main objective is thus to define an appropriate discrete Laplace operator for the mesh.

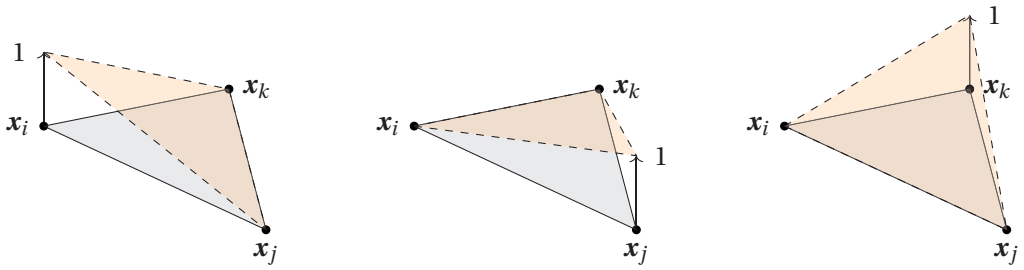


Figure 5.2 – Barycentric basis functions used for interpolation on a triangle.

5.4.2 Discretization of the Laplace operator

In general, the Laplace operator is defined as the divergence of the gradient, $\Delta = \nabla^2 = \nabla \cdot \nabla$. For a function of two parameters $f(x, y)$in Euclidean space, the Laplacian is the sum of second partial derivatives, as shown in equation (5.7).

$$\Delta f = \text{div}\nabla f = \text{div}\begin{pmatrix}\frac{\partial f}{\partial x}\\ \frac{\partial f}{\partial y}\end{pmatrix} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\,. \tag{5.7}$$

In spectral mesh processing, we do not consider the Euclidean space, but the manifold defined by the surface. The Laplace-Beltrami operator generalizes the concept of the Laplace operator to surfaces and, similarly, is defined as $\Delta_{\mathcal{S}} f = \text{div}_{\mathcal{S}} \nabla_{\mathcal{S}} f$ for a function f defined on a manifold surface $\mathcal{S}$. We thus need to define appropriate divergence and gradient operators on manifolds. In the rest of this chapter, we will drop the subscript $\mathcal{S}$ as it should be clear from context that the operators on manifolds are considered.

Discrete gradient

In a triangle mesh, each triangle defines, via its barycentric coordinates, a segment of a piecewise linear surface representation. We start by defining the gradient of a function defined on such a piecewise linear triangle mesh. Such a piecewise linear function f, defined at each mesh vertex as $f(v_i) = f(\boldsymbol{x}_i) = f(\boldsymbol{u}_i) = f_i$, can be interpolated linearly on each triangle $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$ using barycentric basis functions as described in equation (5.8).

$$f(\boldsymbol{u}) = f_i B_i(\boldsymbol{u}) + f_j B_j(\boldsymbol{u}) + f_k B_k(\boldsymbol{u})\,, \tag{5.8}$$

where $\boldsymbol{u} = (u, v)$ are the local coordinates, in the triangle, of the surface point $\boldsymbol{x}$ in a 2D conformal parameterization and $B_{\{i,j,k\}}$ are the barycentric basis functions. These barycentric basis functions are illustrated in figure 5.2. As can be seen in figure 5.2, the gradient of each basis function is orthogonal to the opposite edge of the vertex corresponding to that gradient.

It is given by equation

$$\nabla B_i(\boldsymbol{u}) = \frac{(\boldsymbol{x}_k - \boldsymbol{x}_j)^{\perp}}{2A_T}\,, \tag{5.9}$$

where $\perp$ denotes a counterclockwise rotation by 90 deg in the triangle plane and A_T is the triangle's area.

The gradient of f is given by equation (5.10).

$$\nabla f(\boldsymbol{u}) = f_i \nabla B_i(\boldsymbol{u}) + f_j \nabla B_j(\boldsymbol{u}) + f_k \nabla B_k(\boldsymbol{u})\,. \tag{5.10}$$

We can observe that, since the barycentric basis functions sum up to one everywhere in the triangle, $B_i(\boldsymbol{u}) + B_j(\boldsymbol{u}) + B_k(\boldsymbol{u}) = 1$, their gradient sum to zero, $\nabla B_i(\boldsymbol{u}) + \nabla B_j(\boldsymbol{u}) + \nabla B_k(\boldsymbol{u}) = 0$ and we can rewrite equation (5.10) as equation (5.11).

$$\nabla f(\boldsymbol{u}) = (f_j - f_i)\nabla B_j(\boldsymbol{u}) + (f_k - f_i)\nabla B_k(\boldsymbol{u})\,. \tag{5.11}$$

Thus, the gradient of the piecewise linear function f is given by equation (5.12).

$$\nabla f(\boldsymbol{u}) = (f_j - f_i)\frac{(\boldsymbol{x}_i - \boldsymbol{x}_k)^{\perp}}{2A_T} + (f_k - f_i)\frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)^{\perp}}{2A_T}\,. \tag{5.12}$$

Discrete Laplace-Beltrami operator

There are several ways to discretize the Laplace-Beltrami operator. The two most common discretizations are probably the uniform graph Laplacian, first proposed in [Taubin, 1995], and the cotangent formula. Without going into details, the uniform graph Laplacian only depends on the connectivity of the mesh and, thus, suffers from one major disadvantage: it does not adapt to the spatial distribution of vertices on the surface and, therefore, is not an appropriate discretization for non-uniform meshes. In this thesis, we use the more accurate discretization commonly referred to as the "cotangent formula".

The cotangent formula discretization of the Laplace-Beltrami operator can be derived either using a mixed finite element/finite volume method [Meyer et al., 2002], or using discrete exterior calculus (DEC). Both derivations involve advanced mathematics that are beyond the scope of this thesis. We refer the interested reader to [Lévy and Zhang, 2010] where both derivations are presented.

In this section, we present a simplified derivation, presented in [Botsch et al., 2010], which makes use of the divergence theorem for a vector-valued function $\boldsymbol{f}$ to integrate the divergence of the gradient of a piecewise linear function over a local averaging area Ω_i. The divergence theorem, described in equation (5.13),

$$\int_{\Omega_i} \mathrm{div}\boldsymbol{f}(\boldsymbol{u})\,d\Omega = \int_{\partial\Omega_i} \boldsymbol{f}(\boldsymbol{u}) \cdot \boldsymbol{n}(\boldsymbol{u})\,ds\,, \tag{5.13}$$

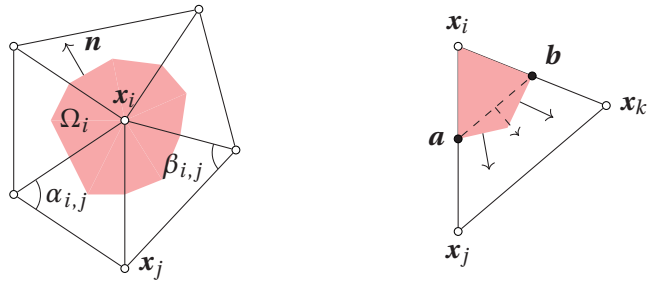


Figure 5.3 – Quantities used in the derivation of the discrete Laplace-Beltrami operator

replaces the integration over the averaging area Ω_i by an integration along its boundary $\partial\Omega_i$, where $\boldsymbol{n}(\boldsymbol{u})$ is the outward pointing normal unit vector of that boundary. Equation (5.14) shows how to apply the divergence theorem to the Laplacian.

$$\int_{\Omega_i} \Delta f(\boldsymbol{u}) d\Omega = \int_{\Omega_i} \mathrm{div} \nabla f(\boldsymbol{u}) d\Omega = \int_{\partial\Omega_i} \nabla f(\boldsymbol{u}) \cdot \boldsymbol{n}(\boldsymbol{u}) ds \,. \tag{5.14}$$

By considering each triangle T separately and the edges' midpoints $\boldsymbol{a}$ and $\boldsymbol{b}$, as illustrated in figure 5.3, we can plug in the definition of the gradient, which is constant within each triangle, given in equation (5.12):

$$\begin{aligned}
\int_{\partial\Omega_i \cap T} \nabla f(\boldsymbol{u}) \cdot \boldsymbol{n}(\boldsymbol{u}) ds &= \nabla f(\boldsymbol{u}) \cdot (\boldsymbol{a} - \boldsymbol{b})^{\perp} \\
&= \frac{1}{2} \nabla f(\boldsymbol{u}) \cdot (\boldsymbol{x}_j - \boldsymbol{x}_k)^{\perp} \\
&= (f_j - f_i) \frac{(\boldsymbol{x}_i - \boldsymbol{x}_k)^{\perp} \cdot (\boldsymbol{x}_j - \boldsymbol{x}_k)^{\perp}}{4\Omega_T} + \\
&\quad (f_k - f_i) \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)^{\perp} \cdot (\boldsymbol{x}_j - \boldsymbol{x}_k)^{\perp}}{4\Omega_T} \,.
\end{aligned} \tag{5.15}$$

Let γ_j, γ_k be the inner triangle angles at vertices v_j, v_k, respectively. Since $A_T = \frac{1}{2} \sin\gamma_j \|\boldsymbol{x}_j - \boldsymbol{x}_i\| \|\boldsymbol{x}_j - \boldsymbol{x}_k\| = \frac{1}{2} \sin\gamma_k \|\boldsymbol{x}_i - \boldsymbol{x}_k\| \|\boldsymbol{x}_j - \boldsymbol{x}_k\|$, $\cos\gamma_j = \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)(\boldsymbol{x}_j - \boldsymbol{x}_k)}{\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \|\boldsymbol{x}_j - \boldsymbol{x}_k\|}$, and $\cos\gamma_k = \frac{(\boldsymbol{x}_i - \boldsymbol{x}_k)(\boldsymbol{x}_j - \boldsymbol{x}_k)}{\|\boldsymbol{x}_i - \boldsymbol{x}_k\| \|\boldsymbol{x}_j - \boldsymbol{x}_k\|}$, equation (5.15) simplifies to equation (5.16).

$$\int_{\partial\Omega_i \cap T} \nabla f(\boldsymbol{u}) \cdot \boldsymbol{n}(\boldsymbol{u}) ds = \frac{1}{2} \left(\cot\gamma_k (f_j - f_i) + \cot\gamma_j (f_k - f_i) \right) \,. \tag{5.16}$$

Thus, integrating over the whole averaging region Ω_i, we obtain

$$\int_{\Omega_i} \Delta f(\boldsymbol{u}) dA = \frac{1}{2} \sum_{v_j \in \mathcal{N}_1(v_i)} \left(\cot\alpha_{i,j} + \cot\beta_{i,j} \right) (f_j - f_i) \,, \tag{5.17}$$

where $\mathcal{N}_1(v_i)$ is the one-ring neighborhood of vertex v_i and the angles $\alpha_{i,j}$ and $\beta_{i,j}$ are illustrated in figure 5.3. The discrete average of the Laplace-Beltrami operator of a function f

at a vertex v_i over the region Ω_i is finally described in equation (5.18).

$$\Delta f(v_i) = \frac{1}{2\Omega_i} \sum_{v_j \in \mathcal{N}_1(v_i)} \left(\cot \alpha_{i,j} + \cot \beta_{i,j}\right)(f_j - f_i) \,. \tag{5.18}$$

This allows to discretized the Laplace-Beltrami operator Δf at each mesh vertex v_i by a linear combination of the function values at v_i and at its one-ring neighbors v_j:

$$\Delta f(v_i) = w_i \sum_{v_j \in \mathcal{N}_1(v_i)} w_{i,j} \left(f(v_j) - f(v_i)\right) . \tag{5.19}$$

Stacking the function values $f(v_i)$ and Laplacians $\Delta f(v_i)$ for all n vertices allows to write the discrete Laplacian of the mesh in matrix notation, as described in equation (5.20).

$$\begin{pmatrix} \Delta f(v_1) \\ \vdots \\ \Delta f(v_n) \end{pmatrix} = \underbrace{D^{-1}Q}_{L} \begin{pmatrix} f(v_1) \\ \vdots \\ f(v_n) \end{pmatrix} . \tag{5.20}$$

$D = \text{diag}(w_1, \ldots, w_n)$ is a diagonal matrix of vertex weights $w_i = \Omega_i$ and Q is a symmetric matrix of edge weights.

$$Q_{i,j} = \begin{cases} \frac{1}{2}\left(\cot(\alpha_{i,j}) + \cot(\beta_{i,j})\right) , & \text{when } v_j \in \mathcal{N}_1(v_i) \,, \\ -\sum_{v_k \in \mathcal{N}_1(v_i)} Q_{i,k} \,, & \text{when } i = j \,, \\ 0 \,, & \text{otherwise.} \end{cases} \tag{5.21}$$

5.4.3 Band-by-band eigendecomposition

The eigen-decomposition of the discrete Laplace operator is obtained by solving equation (5.22) for the eigenvectors $\boldsymbol{h}_k$ and eigenvalues λ_k.

$$-L\boldsymbol{h}^k = \lambda^k \boldsymbol{h}^k \tag{5.22}$$

The Laplace operator matrix L can be large, depending on the number of vertices of the mesh n but is sparse. To compute the solutions of a large sparse eigenproblem, several iterative algorithms exist. The publicly available library ARPACK[4] provides an efficient implementation of the implicit restarted Arnoldi method for iteratively solving large-scale sparse eigenvalue problems.

There are two main obstacles to the computation of the eigen-decomposition of a large discrete Laplacian matrix L:

- Iterative solvers perform better at computing high frequencies, i.e. eigenvectors associ-

[4] A C++ interface to the ARPACK Fortran package is available at https://github.com/m-reuter/arpackpp

ated with high eigenvalues, but we are interested mainly in low frequencies.

- The computation time is superlinear with the number of eigenpairs and we need to compute a large number of eigenvectors.

Both issues can be addressed by using the band-by-band algorithm proposed in [Vallet and Lévy, 2008]. It takes advantage of the *Shift-Invert* spectral transform. First, the spectrum is shifted by λ_S by replacing L with $L - \lambda_S \mathrm{Id}$ and then swapped by inverting this matrix as $L^{SI} = (L - \lambda_S \mathrm{Id})^{-1}$. This allows to define a new eigenproblem, as desribed in equation (5.23), which have the same eigenvectors as the original one and which eigenvalues are related to the original ones by $\lambda_k = \lambda_S + \frac{1}{\mu_k}$.

$$-L^{SI}\boldsymbol{h}^k = \mu_k \boldsymbol{h}^k \tag{5.23}$$

It thus becomes possible to apply an iterative solver, which will return the high end of the spectrum efficiently, i.e. the largest μ_k, which corresponds to a band of eigenvalues centered around λ_S. The band-by-band algorithm 2 splits the computation into multiple bands and obtain a computation time that is linear in the number of computed eigenpairs.

Algorithm 2 Band-by-band algorithm

$\lambda_S \leftarrow 0$; $\lambda_{last} \leftarrow 0$
while $\lambda_{last} < \omega_m^2$ **do**
    compute an inverse L^{SI} of $(L - \lambda_S \mathrm{Id})$
    find the 50 first eigenpairs $(\boldsymbol{h}^k, \mu_k)$ of L^{SI}
    for $k = 1$ to 50 **do**
        $\lambda_k \leftarrow \lambda_S + \frac{1}{\mu_k}$
        if $\lambda_k > \lambda_{last}$ **then**
            write $(\boldsymbol{h}^k, \lambda_k)$
        end if
    end for
    $\lambda_S \leftarrow \max(\lambda_k) + 0.4(\max(\lambda_k) - \min(\lambda_k))$
    $\lambda_{last} \leftarrow \max(\lambda_k)$
end while

5.5 Conclusion

In this chapter, we first reviewed existing databases of 3D facial scans, both unregistered and registered, and nonrigid registration methods. A large number of unregistered databases have been proposed, since the 2000s, for different applications such as face recognition and identification, a early popular research topic in facial image analysis, or head pose estimation. In the scope of this thesis, we would like to emphasize that, even though these databases cannot be used directly to build 3D models of the face because their scans do not share a common parameterization, they can be used with an appropriate 3D nonrigid registration

method, such as the one proposed in chapter 6. This represents a large potential source of 3D facial scans, which could augment registered databases, mostly with respect to neutral facial expression scans.

In the second section of this chapter, we provided a introduction to the Kinect Fusion algorithm, which allows to fuse noisy and incomplete depth maps from different points of view into a high quality 3D reconstruction of the scene. We have used this algorithm to collect all the data of the EPFL3DFace database, which we introduce in the next chapter 6.

Finally, we provided an overview of some important aspects of spectral geometry processing. We make extensive use of spectral geometry processing in the next chapter of this thesis, chapter 6. The use of these methods in computer graphics is relatively new and despite some very good tutorials in the main conferences in the field [Lévy and Zhang, 2010, Chang et al., 2010], resources that introduce these methods in an accessible way are scarce.

6 Spectral nonrigid registration

6.1 Introduction

In this chapter, we describe a 3D nonrigid registration method based on a spectral embedding of the source and an implicit representation of the target. In order to build a new 3D face model, which is able to describe variations due to specific facial expressions as well as a specific population, we recorded a new database of facial scans and applied the proposed method to register these scans.

In this chapter, we first propose to compute a spectral embedding of the source and use that representation to constrain the possible deformations. Deforming the source in the spectral domain allows choosing which frequency band to focus on, depending on required properties. In our nonrigid registration pipeline, we propose to embed the template in the spectral domain using a *manifold harmonics transform* (MHT) [Vallet and Lévy, 2008] and use this embedding as a surface deformation model. Indeed, by optimizing over the parameters corresponding to lower frequencies, we enforce the deformation to be smooth. Moreover, depending on the number of frequencies M_{freq} chosen, the number of parameters to optimize, $3 \times M_{\text{freq}}$, is much smaller than in the case of *per-vertex affine transform*, $12 \times N_{\text{vert}}$ as $M_{\text{freq}} < N_{\text{vert}}$. As an example, in our experiments, the template has $N_{\text{vert}} = 11510$ vertices. That would results in 138'120 parameters to optimize in a *per-vertex affine transform* model but our spectral embedding uses 500 basis functions, resulting in 1500 parameters to optimize in our transformation model, thus reducing the number of parameters by a factor 92.

A second keypoint of our method is the implicit surface representation [Ohtake et al., 2003] of the target three dimensional (3D) scans in order to overcome the problem of point correspondence. We propose to represent the target as an implicit surface in order to avoid computing correspondences, when evaluating the distance between the source and the target and the gradient of that distance. By representing the target as an analytical implicit surface, defined as the zero level-set of a squared distance function, the distance of any point to the surface is obtained by evaluating the value of the implicit function at that point. Moreover, when computing the implicit surface representation, the implicit function can approximate the

original scan, rather than interpolate it, thus effectively removing noise and filling holes.

Finally, we contribute to the availability of more 3D facial surfaces by introducing EPFL3DFace, a new database consisting of 120 subjects performing 35 expressions. We show that the subspace spanned by our 120 subjects, among which 87% are Caucasian, extends the subspace spanned by the subjects from FaceWarehouse [Chen Cao et al., 2014], another publicly available database of fully registered 3D facial scans including a variety of facial expressions.

Establishing correspondences from one surface to another has been investigated in several fields and under different names such as *nonrigid registration*, *alignment*, *matching*, *mesh morphing*, *cross-parameterization* or *correspondence estimation*. A few of the most relevant methods are discussed hereafter and we refer the reader to the book of Bronstein et al. [Bronstein et al., 2008] or the surveys of Van Kaick et al. [van Kaick et al., 2011] and Tam et al. [Tam et al., 2013] for more exhaustive reviews of the different methods.

In the remaining of this chapter, we describe the new nonrigid registration method that we propose in section 6.2. We then introduce EPFL3DFace, our database of 3D facial expressions in section 6.3 and present results achieved by the proposed method on the new database in section 6.4. Finally section 6.5 summarizes the contributions of this chapter and discusses a few directions for future work.

6.2 Methods

The complete alignment pipeline is composed of the following steps, described in detail in the following subsections: first, the template is rigidly aligned to the target such that both surfaces share the same scale, position and orientation in space. This initial rigid alignment is described in subsection 6.2.1. The different parts of the nonrigid registration are then described in subsection 6.2.2: the similarity measure using implicit surface representation, the transformation model using Manifold Harmonics Transform (MHT), and the complete objective function and optimization process.

6.2.1 Initial rigid 3D registration

Our scans are, in general, not rigidly aligned with the template. Before being able to nonrigidly align the source to the target, it is essential to compensate for unknown rigid transformations such as scale, translation and rotation.

3D feature points, or landmarks, are used to compute the rigid transform between the source and the target such that the source is rigidly aligned to the target. First, 68 landmarks are manually annotated on the source. Note that this is done only once as the sources used for each expression are already registered.

Then, similar to the approach used in the LSFM [Booth et al., 2016], we automatically detect

the same 68 landmarks on each target. An image is first generated by projecting the 3D surface on the image plane of a frontal virtual camera. We then detect the landmarks on this image using a state-of-the-art facial feature detection algorithm [Qu et al., 2015] based on the supervised descent method (SDM) [Xiong and Torre, 2014]. In order to get the 3D positions of the landmarks on the target, we back-project the two dimensional (2D) positions of the landmarks with the known projection matrix of the virtual camera and intersect these rays with the 3D surface. The landmarks on the jaw are often less precisely located on the 3D surface due to the fact that the back-projected rays are almost tangential to the surface and thus a small imprecision in 2D becomes a large error in the intersection. For that reason, we discard these when computing the rigid transform.

Finally, the rigid transform, *i.e.* the translation and rotation between the two sets of 3D landmarks is computed as a weighted least-squares problem using a singular value decomposition (SVD) [Arun et al., 1987, Sorkine, 2009]. The scaling factor between the two sets is retrieved as well. The scaling, translation, and rotation are applied to the source and the resulting shape is used for the nonrigid registration described in the next section.

6.2.2 Nonrigid 3D registration

Each scanned 3D facial surface needs to be re-parametrized into a consistent form, where the number of vertices, the triangulation, and the anatomical meaning of each vertex are consistent across all surfaces. The general strategy is for each scan to deform a rigidly aligned template, the *source*, $\mathcal{S}$ such that it matches the scan or *target surface*, $\mathcal{T}$. The deformation model, which ensures a meaningful deformation, is denoted by χ and the quality of the match is measured by a similarity measure.

$$\mathcal{S} = \{\boldsymbol{p}_i | i = 1, ..., N^S\} \overset{\chi}{\mapsto} \mathcal{T} = \{\boldsymbol{q}_i | i = 1, ..., N^T\} . \tag{6.1}$$

This dense correspondence problem is referred to as nonrigid registration and is defined by three main elements:

- a similarity measure, dependent on the representations of the *source* $\mathcal{S}$ and the *target* $\mathcal{T}$,
- a transformation model χ, which describes allowed deformations of the source, and
- an objective function, which combines the similarity measure and the transformation model and is optimized with a numerical optimizer.

In the next subsections, we will detail each of these three elements.

Similarity measure with implicit surface representation

In classical nonrigid iterative closest point (NICP) approaches, correspondences need to be computed in order to be able to evaluate the distance between the source and the target. These are unknowns as this is precisely what we are looking for in the first place. Several iterative approaches have been proposed based on spatial proximity of points, either using a *point-to-point* or a *point-to-plane* distance and looking for correspondences from the source to the target or the opposite, or a combination of both [Weise et al., 2009, Mitra et al., 2004]. The correspondence problem gets even more complicated, when the quality of one or both surfaces is low. In particular, holes and noisy parts in the target further complicate the search for correspondence.

We propose to use an implicit surface representation for the target in order to avoid having to estimate correspondences. The surface is then implicitly represented as the zero level-set of a distance function $\mathfrak{d} : \mathbb{R}^3 \mapsto \mathbb{R}$. Choosing carefully that function allows to approximate the input surface rather than interpolate it, thus smoothing it and filling holes. In addition, desirable properties of an implicit surface reconstruction method include speed and low memory overhead.

As the value of the function is the signed distance to the surface, evaluating a distance between the source and the target can be achieved by simply summing the squared value of the implicit function at each vertex of the source, as described in equation (6.2). This does not require searching for correspondences.

$$\mathrm{dist}^2(\mathcal{S},\mathcal{T}) = \sum_i \mathfrak{d}\left(\boldsymbol{p}_i\right)^2 . \tag{6.2}$$

Multilevel partition of unity (MPU) implicits provide fast, accurate, and adaptive reconstructions of complex shapes [Ohtake et al., 2003]. The main advantage of MPU is to define approximants locally, thus avoiding the overhead of a global support, and integrate them together by weighting each of them. The local approximants Q_i, in each cell of the OCtree are blent with smooth, local weights w_i that sum up to one everywhere on the domain, as described by equation (6.3).

$$f(\boldsymbol{x}) \approx \sum_i \phi_i(\boldsymbol{x}) Q_i(\boldsymbol{x}) \qquad \text{with} \qquad \sum_i \phi_i \equiv 1 , \tag{6.3}$$

where $f(\boldsymbol{x})$ is the function to approximate and ϕ_i is the partition of unity function for a given cell of the OCtree. The partition of unity functions are described by equation (6.4).

$$\phi_i(\boldsymbol{x}) = \frac{w_i(\boldsymbol{x})}{\sum_{j=1}^{n} w_j(\boldsymbol{x})} . \tag{6.4}$$

Following the original method, we use the quadratic B-spline $b(t)$ to generate weight functions

$$w_i(\boldsymbol{x}) = b(\frac{3|\boldsymbol{x}-\boldsymbol{c}_i|}{2R_i} + \frac{3}{2}) \,, \tag{6.5}$$

centered at $\boldsymbol{c}_i$ and with a spherical support of radius R_i.

MPU uses a hierarchical structure to adaptively divide the region of space containing the input set of shape vertices. We use an OCtree structure, starting from the bounding cube of the shape and computing an approximation of the points enclosed in a sphere of radius R. The radius of the sphere is proportional to the main diagonal d of the current cell $R = \alpha d$. When the computed local max-norm approximation error ϵ is greater than a user-specified threshold ϵ_0, the cell is subdivided and the process is repeated. This allows the OCtree to adapt to the relation between local shape complexity and desired accuracy.

If the initial sphere does not contain enough points to compute the approximation, the radius is iteratively increased until the sphere contains a user-defined minimum number of points N_{min}. In that case, the cell is not further subdivided, independently of the approximation error and unlike the original method in which the initial sphere needs to be empty to stop the subdivision. The local max-norm approximation error ϵ is estimated according to the Taubin distance [Taubin, 1995] and is given by equation (6.6).

$$\epsilon = \max_{|\boldsymbol{p}_i-\boldsymbol{c}|<R} |Q(\boldsymbol{p}_i)|/|\nabla Q(\boldsymbol{p}_i)| \,. \tag{6.6}$$

The choice of the approximants allows to address different scenarios: locally planar surfaces, surfaces with sharp edges, etc., as emphasized in [Ohtake et al., 2003]. Following the original method, we implemented the bivariate quadratic polynomial and the general quadric approximants. To give an intuition, the bivariate quadratic polynomial is best suited to approximate local smooth patches, and the general quadric provides consistent approximations on larger parts of the surface which might contain more than one sheet.

In practice, the surfaces we are implicitly representing, our scans, are mainly composed of local smooth patches in the region of interest, the face region, and noisy boundaries. Therefore, we only use the bivariate quadratic polynomial approximant. This and the choice of N_{min} have shown to be critical, when implicitly representing the scans from our database as explained in section 6.3.

Transformation model

When deforming the source toward the target, the transformation model defines the possible transformations of the source in order to avoid overfitting, prohibit arbitrary deformations, and favor reasonable ones and reduce the dimensionality of the problem. Intuitively, coarse, global deformations should be applied first and then refined with fine, local deformations. In general, smoothness should also be preserved.

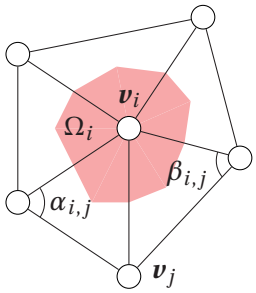


Figure 6.1 – Angles and local averaging area, Ω_i, used in the discrete Laplace-Beltrami operator

Per-vertex displacements are thus modeled using spectral tools [Lévy and Zhang, 2010]. They offer an intuitive control over deformations where coarse, global deformations are embedded in the low frequencies and fine, localized deformations in the high frequencies. By selecting a number of lower frequencies $m \ll n$, the number of vertices in the source, the dimensions of the optimization problem are reduced. Moreover, the built-in smoothness of the low frequencies helps to avoid overfitting.

The Laplacian framework and differential representations allow to describe surface meshes through their differential properties. As a generalization of Fourier analysis the *Manifold Harmonics Basis* (MHB) and corresponding Manifold Harmonics Transform (MHT) introduced in [Vallet and Lévy, 2008] provide a re-parametrization tool which allows us to represent a mesh with potentially fewer coefficients and more interestingly to constrain the deformation of the mesh, when changing the coefficients in ways that preserve the smoothness of the mesh.

Manifold harmonics are defined as the eigenfunctions of the discrete Laplace operator. The basis vectors of the MHT are thus the eigenvectors $\boldsymbol{h}^k$ of the discrete Laplacian as described in equation (6.7).

$$\boldsymbol{h}^k = [H_1^k, ..., H_n^k] \quad \text{satisfies} \quad -Q\boldsymbol{h}^k = \lambda D\boldsymbol{h}^k \,. \tag{6.7}$$

The matrix Q is called the *stiffness matrix* and is defined by the *cotangent formula*:

$$Q_{i,j} = \begin{cases} \frac{1}{2}\left(\cot(\alpha_{i,j}) + \cot(\beta_{i,j})\right) & \text{when } i \neq j \\ -\sum_k Q_{i,k} & \text{when } i = j \,. \end{cases} \tag{6.8}$$

where the angles $\alpha_{i,j}$ and $\beta_{i,j}$ are illustrated in figure 6.1.

The diagonal matrix D is called the *lumped mass matrix* and is defined by:

$$D_{i,i} = \sum_{t \in St(i)} \Omega_t \,, \tag{6.9}$$

where $St(i)$ denotes the set of triangles incident to i and Ω_t the local averaging area of triangle

t. In our case, we use the barycentric cell as local averaging area. The barycentric cell connects the triangle barycenter with the edges' midpoints. The eigendecomposition of the discrete Laplacian described by equation (6.7) is computed using the band-by-band algorithm described in [Vallet and Lévy, 2008], which takes advantage of the *Shift-Invert* spectral transform.

To compute the transform of the function x from geometric space to frequency space, x is projected onto the manifold harmonics basis through the inner product. The MHT of x is a vector $[\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m]$ given by equation (6.10).

$$\tilde{x}_k = < x, H^k > = \boldsymbol{x}^T D \boldsymbol{h}^k = \sum_{i=1}^{n} x_i D_{i,i} H_i^k \ . \tag{6.10}$$

The inner product contains D in order to ensure orthogonality of the basis, as the Laplacian is not symmetric, due to the weights $D_{i,i}$ which scale the lines of Q.

The inverse transform, to map the function $\tilde{x}$ in frequency space into its geometric space is given by equation (6.11).

$$x_i = \sum_{k=1}^{m} \tilde{x}_k H_i^k \ . \tag{6.11}$$

H is a basis containing the spectral modes of variation of the shape. We thus represent a new shape as the original source shape $\bar{\boldsymbol{p}}$ and a linear combination of spectral deformations, as described in equation (6.12).

$$\boldsymbol{p}(\boldsymbol{\alpha}) = \bar{\boldsymbol{p}} + H\boldsymbol{\alpha} \ , \tag{6.12}$$

where $\boldsymbol{\alpha}$ is a vector of spectral coefficients. Setting $\boldsymbol{\alpha}$ to zero yields the initial shape, without deformation.

Furthermore, as described in section 6.2.1, the source has been rigidly aligned to the target beforehand. Nevertheless, as pointed out by Blanz et al. [Blanz et al., 2004], the result of this rigid pre-alignment is sub-optimal, since the optimal rigid alignment depends on the source after deformation. Thus we need to include translation and rotation in the transformation model. Translation is included in the first spectral basis, which is a constant vector, and we include a linearized rotation similarly to [Blanz et al., 2004] as described in equation (6.13).

$$\begin{aligned} \boldsymbol{R}\boldsymbol{v} &\approx c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + \boldsymbol{v} \\ \boldsymbol{s}_\gamma &= (-y_1, x_1, 0, -y_2, x_2, 0, ...)^T \\ \boldsymbol{s}_\theta &= (0, -z_1, y_1, 0, -z_2, y_2, ...)^T \\ \boldsymbol{s}_\phi &= (z_1, 0, -x_1, z_2, 0, -x_2, ...)^T \ . \end{aligned} \tag{6.13}$$

The complete transformation model is thus given by equation (6.14).

$$\boldsymbol{p}(c_\gamma, c_\theta, c_\phi, \boldsymbol{\alpha}) = \bar{\boldsymbol{p}} + c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + H\boldsymbol{\alpha} \,. \tag{6.14}$$

Objective function

Combining the transformation model and the implicit surface distance measure, we can evaluate the similarity between the deformed source and the target for a given set of parameters $\boldsymbol{\alpha}$, c_γ, c_θ, c_ϕ. We define the data fitting term E_{data} of our objective function as in equation (6.15).

$$E_{data} = \mathfrak{d}^t \left(\bar{\boldsymbol{p}} + c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + H\boldsymbol{\alpha} \right) \,. \tag{6.15}$$

We noticed that, due to the relatively low accuracy of the Kinect, the eye regions often do not contain enough details to correctly align the eyes. This causes the eyes of the source to slide on the flat region around the eyes of the target surface, ending in incorrect positions. To further constrain the eye regions, we use 3D landmarks around the eyes. On the target, these landmarks are detected with high accuracy during the rigid alignment step, whereas on the source, they have been manually annotated. The landmarks detection and annotation process is detailed in section 6.2.1. In order to constrain the eye regions, we add a term to the objective function penalizing large distances between the landmarks on the source and the corresponding landmarks on the target. This term is defined in equation (6.16).

$$E_l = \sum_{i=1}^{n_l} \| \hat{p}_i - \hat{q}_i \|_2^2 \,, \tag{6.16}$$

where n_l is the number of landmarks, $\hat{p}_i$ are the landmarks on the source and $\hat{q}_i$ are the landmarks on the target.

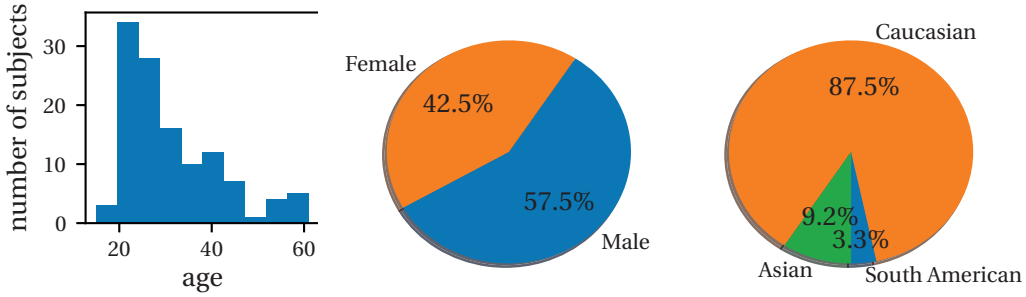
As discussed in section 6.2.2, we want to favor low frequencies over high frequencies, thus we add a regularization term E_b to penalize higher bending of the deformation. This regularization term is defined in equation (6.17).

$$E_b = \| \Lambda_H \boldsymbol{\alpha} \|_2^2 \,, \tag{6.17}$$

where Λ_H is a diagonal matrix of eigenvalues corresponding to the spectral bases.

A second regularization term E_m penalizes the magnitude of the deformation, as defined in equation (6.18).

$$E_m = \| \boldsymbol{\alpha} \|_2^2 \,. \tag{6.18}$$



(a) (b) (c)

Figure 6.2 – (a) Age, (b) gender and (c) ethnicity distributions of the subjects included in the database

The complete objective function is given in equation (6.19).

$$E = E_{data} + \beta_0 E_l + \beta_1 E_b + \beta_2 E_m \,. \tag{6.19}$$

We use a gradient descent solver to minimize E. In our experiments, we chose $\beta_0 = 1e^{-4}$, $\beta_1 = 2e^{-3}$ and $\beta_2 = 2e^{-4}$ empirically.

6.3 EPFL3DFace database

We have collected EPFL3DFace, a new 3D facial expressions database, for the study. The 3D facial surfaces have been nonrigidly registered with the method presented such that they are all in dense correspondence. This allows the use of EPFL3DFace database to train a 3D statistical model of the face, for example a morphable model, a multilinear model or a blendshape model, for a large variety of applications, such as but not limited to facial expression recognition, visual speech recognition, morphological analysis of the face, etc.

We recorded 120 subjects performing 35 facial expressions, while sitting still on a rotating chair. The subjects were facing a Microsoft Kinect® for Windows v.1 at a distance of 50-70 cm. A screen in front of the subjects was displaying instructions on how to perform each expression with visual examples. At the same time, an operator was explaining and demonstrating how to perform the expression. Each subject had to perform each expression and stay perfectly still, while the operator was rotating the chair at an angle of $\pm 60° - 90°$. This operation took approximately 15 seconds on average.

Figure 6.2 shows the age, gender, and ethnicity distributions of the subjects included in EPFL3DFace database. In general, the population is slightly biased towards young men, as the subjects were recruited mainly in the electrical engineering department of the university.

(a) (b) (c)

Figure 6.3 – Examples of scans from the database: (a) jaw forward (AU29), (b) viseme /uh/, (c) surprise

With 43% women and 57% men, the gender distribution is still reasonably well balanced. The ethnicity is strongly biased towards Caucasian, with 87% of the subjects included in the database. This is a wanted feature of the database making it complementary to FaceWarehouse [Chen Cao et al., 2014], which mainly contains Asian subjects. We discuss this aspect in more detail in section 6.4.2. We also recorded the country of origin and the mother-tongue of the subjects.

We recorded each subject with a neutral facial expression, with the eyes open, and then instructed them to perform different facial expressions. These include prototypical expressions: anger, sadness, surprise, fear, disgust, happiness, and variants: anger with mouth slightly open, sad surprise and grin. They also include specific action units (AU): closed eyes (AU43), mouth open (AU25), brow lower (AU04), brow raiser (AU01), jaw left and right (AU30), jaw forward (AU29), mouth left and right, dimples (AU14), chin raiser (AU17), lips funneler (AU22), lips puckerer (AU18), lips roll (AU28), and cheek blow (AU33). Nine visemes are also included representing the following phonemes /ah/, /uh/, /axr/, /eh/, /l/, /m/, /n/, /f/, /iy/ and one extreme facial movement: biting their own top lip.

In order to generate a smooth and low-noise 3D mesh from noisy and incomplete depth maps, we aggregated multiple depth maps from different view points in order to construct a full view of the face for each expression and subject by using the Kinect Fusion algorithm[1] [Newcombe et al., 2011, Izadi et al., 2011]. Thus, a 3D facial surface was obtained for each expression of each subject. Figure 6.3 shows three examples of obtained scans.

[1] A lightweight, reworked and optimized version of KinFu, originally shared in PCL in 2011, is available on https://github.com/Nerei/kinfu_remake

6.3.1 Nonrigid alignment of the database scans

As mentioned in chapter 5, in order to allow for statistical modeling of the faces in EPFL3DFace database, these need to be put in dense correspondence. We nonrigidly align all the scans in the database such that all the expressions of all the subjects share a common parametrization using the method described in section 6.2. This allows for statistical modeling of the variations due both to the identity and the expression.

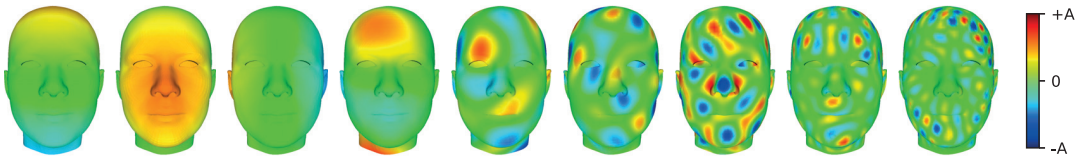
Our method is based on the deformation of a single template shape. The advantage of not requiring a full statistical shape model (see chap. 5) but only a static template comes at the price of a larger sensitivity to the initialization. This implies that in order to converge to the target, the initial template to be deformed should be close already. Since we observed that the 3D shape of the face varies significantly due to changes in facial expressions, we decided to use a separate template for each facial expression.

We take advantage of the FaceWarehouse [Chen Cao et al., 2014] database and compute one mean shape for each expression. For each expression in EPFL3DFace, we select as template the FaceWarehouse mean shape closest to that expression. Some expressions have direct correspondences in both databases as the set of expressions from FaceWarehouse is included in EPFL3DFace. For the remaining expressions in EPFL3DFace, we manually selected the closest corresponding expression in FaceWarehouse.

An important advantage of using different templates for each expression is that we do not need to perform any kind of expression transfer. Indeed, the templates of all expressions are already registered together. After registration, the scans of different expressions are in dense correspondence, since the templates used for registration are in dense correspondence.

In practice, we do not evaluate all the vertices of the source in the implicit function of equation (6.2), but only the vertices lying on the face. Due to the fact that we use the closest expression mean shape of FaceWarehouse as the source for each expression in EPFL3DFace, the topologies of the source and the targets are very different. FaceWarehouse mean shapes are closed surfaces, homeomorphic to a sphere, whereas the scans are bounded surfaces, homeomorphic to a plane. Moreover, the scans of the head are only partial and information is missing on the top and the back of the head. That is not the case with the FaceWarehouse shapes. Trying to align all the vertices of such shapes onto our scans would not be reasonable as they do not share the same topologies and do not contain the same information even though there is an overlap. Thus, we define the set of landmarks lying on the face to use in the implicit distance computation defined in equation (6.2). Note that the deformation is still applied to the whole shape. In summary, the whole source shape is deformed such that the distance between vertices on the face and the target is minimized.

This nonrigid alignment process is repeated for each scan in the database, resulting in a database of registered 3D surfaces of 120 subjects, performing 35 different expressions and facial movements. This database is available to the research community upon request.



(a) b_1 (b) b_2 (c) b_3 (d) b_9 (e) b_{49} (f) b_{99} (g) b_{199} (h) b_{349} (i) b_{499}

Figure 6.4 – Visualization of some of the spectral bases b_i. The amplitude of the deformation for each vertex is normalized over the first 500 bases where $-A$ is the maximum deformation amplitude towards the inside of the surface and $+A$ the maximum towards the outside of the surface.

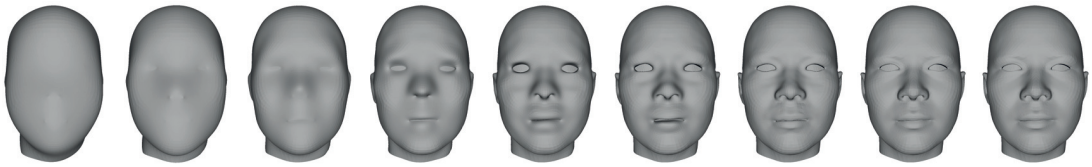
6.4 Results

In this section we discuss qualitative results obtained with the proposed method on the collected database. In subsection 6.4.1, we show a few of the manifold harmonic bases that are used to constrain the nonrigid deformation as well as different reconstructions obtained with a varying number of bases and discuss the influence of the number of bases. In subsection 6.4.2, we then show visual results and compare the obtained deformed shapes with their corresponding targets. We also provide detailed visualization of the spectral deformation process and analyze the evolution of the different terms in the objective function. Finally in subsection 6.4.3, we visualize the manifolds of shapes and compare these manifolds between an existing database, FaceWarehouse, and our new database.

The lack of ground truth is the main obstacle to a quantitative validation of the method. Indeed, as it is a dense registration problem, the locations of each and every landmarks of the source would need to be manually annotated. Depending on the number of vertices in the source, this represents several thousands of 3D locations for each 3D face scan. Moreover, this problem is largely under-constrained. Ultimately, the topology and geometry of the target is transferred to the source, but these are not uniquely defined by the 3D locations of the vertices. As an example, moving one vertex of the source along the surface of the target does not necessarily change the quality of the alignment.

6.4.1 Spectral basis visualization

Figure 6.4 shows the first 3 bases and a few other bases corresponding to higher frequencies. Note that basis 0 is constant and is not depicted in the figure. As expected, spectral bases corresponding to lower frequencies show smoother deformations of the surface, whereas higher frequencies provide more localized deformations. The choice of the number of bases to consider in the deformation model is thus guided by the level of details at which the deformation is expected to fit. A very important consideration is that this resolution is only the resolution of the deformation and not the resolution of the obtained mesh. Indeed, the spectral content of all other frequencies outside the frequency band considered in the deformation



(a) $N = 50$ $E = 78.28\%$ (b) $N = 100$ $E = 82.13\%$ (c) $N = 200$ $E = 85.66\%$ (d) $N = 500$ $E = 90.02\%$ (e) $N = 1000$ $E = 92.91\%$ (f) $N = 2000$ $E = 95.52\%$ (g) $N = 5000$ $E = 98.74\%$ (h) $N = 10000$ $E = 99.95\%$ (i) original

Figure 6.5 – (a)-(h) Reconstructions of the FaceWarehouse neutral mean shape using the first N bases, keeping E percent of the energy. (i) The original shape.

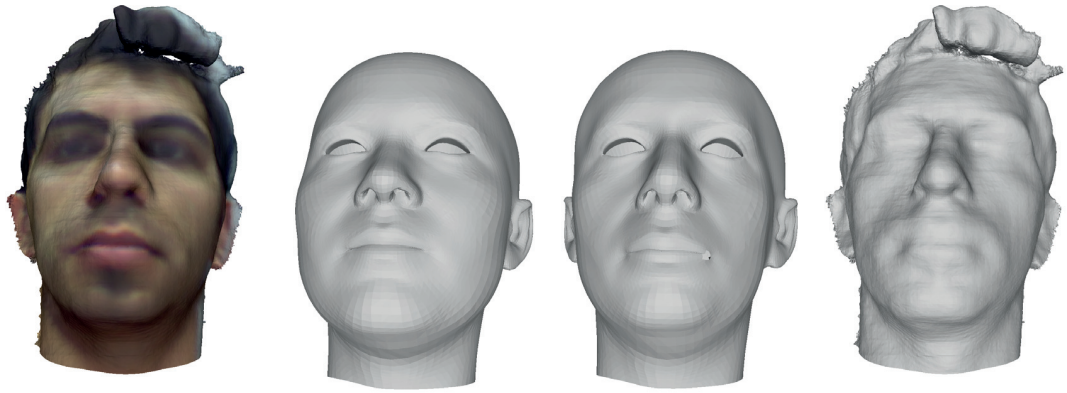
model is retrieved from the source shape $\bar{\boldsymbol{p}}$ in equation (6.12).

In order to get a better intuition of the resolution of the deformation, figure 6.5 shows different reconstructions of the FaceWarehouse [Chen Cao et al., 2014] mean shape with neutral expression. These were obtained by computing the MHT, transforming the shape into the spectral domain, setting all the spectral coefficients $\tilde{x}$ to zero except the first N coefficients and taking the inverse transform to return to the spatial domain. In short, the source has been filtered with a low-pass filter, whose cut-off frequency varies with the number of bases kept.

Experimentally, we found that keeping the first 500 bases is a reasonable trade-off between the resolution of the deformation and the compactness of the deformation model. The energy of the template that is kept in these 500 bases corresponds to 90.02% of the total energy. With 500 bases, the resolution of the deformation is sufficient to deform shapes with a given expression toward the scans representing the same expression or close ones on subjects with different identities, as explained in section 6.4.2.

6.4.2 Spectral alignment

We present the results of the complete nonrigid alignment process on a neutral scan of the EPFL3DFace database in figure 6.6. Figure 6.6a shows the clean and normalized color scan from the database. The corresponding mean shape from FaceWarehouse is then rigidly aligned to the scan using 3D landmarks, as detailed in section 6.2.1. In that case, the source is the neutral expression mean shape. The resulting scaled, translated, and rotated mean shape is shown in figure 6.6b. That rigidly aligned source is then nonrigidly deformed following the method described in section 6.2.2 and the result is shown in figure 6.6c. For better visual comparison, the target is shown again, without texture, in figure 6.6d. It should be noted that even though the rigid alignment does not retrieve the exact pose of the target, as shown by a comparison of the head poses between figures 6.6a and 6.6b, this misalignment is corrected during the nonrigid alignment by the linearized rotation term of the transformation model described in equation (6.13). Figure 6.7 presents additional results on two subjects performing seven other expressions or facial movements.



(a) (b) (c) (d)

Figure 6.6 – Alignment results. (a) Color target (b) Rigidly aligned source (c) Result of the nonrigid alignment (d) Target without color (for better comparison).

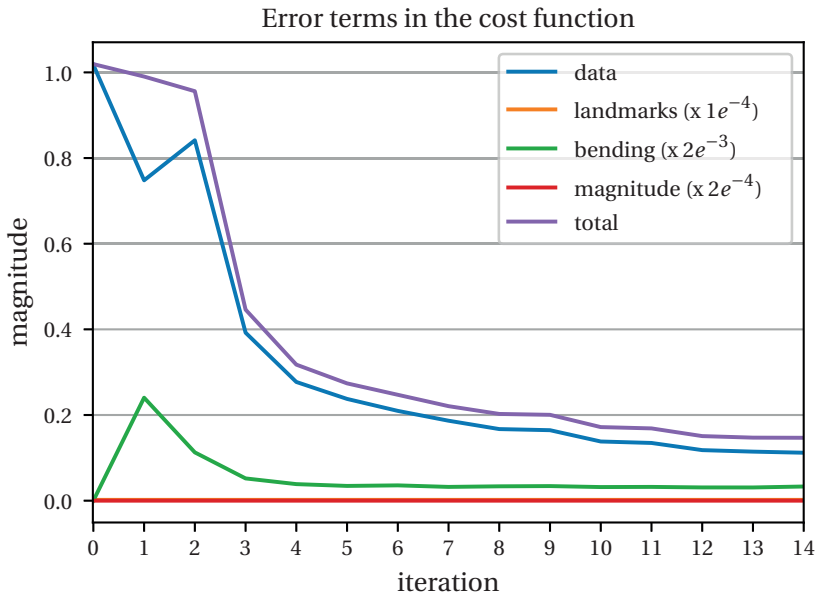
In order to get a better understanding of the spectral deformation process, figure 6.8 shows the evolution of the different terms in the objective function as well as corresponding shapes, magnitudes of deformation, and distances to the target for a few steps of the optimization. Overall, the data term and the sum of all terms decrease with the number of iterations and seem to have converged at the end of the optimization process. The role of the bending regularization term is clear in the first steps of the optimization, where it prevents extreme, non-realistic deformations to dominate as seen in iteration 1 in figure 6.8a.

6.4.3 Facial manifold visualization

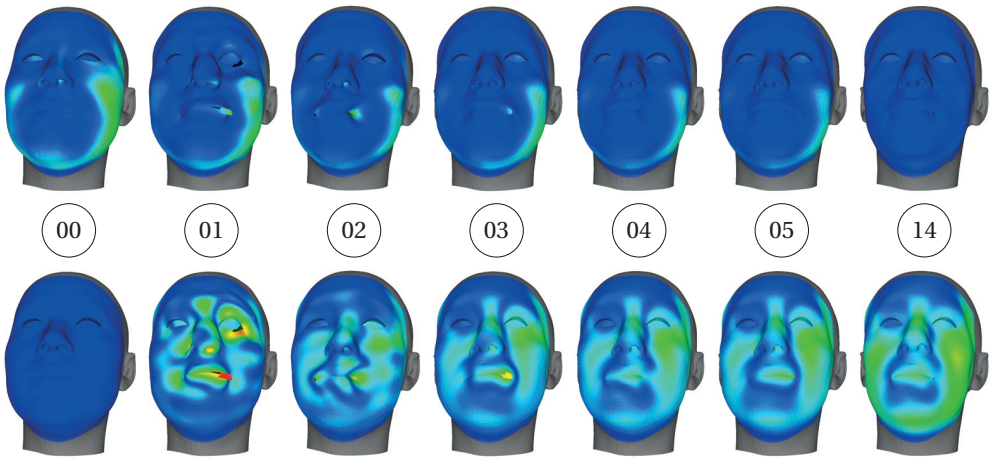
In order to validate the intuition that training 3D face models using scans of people from different populations yields different manifolds, we visualize the manifold of scans from the FaceWarehouse database as well as EPFL3DFace using t-SNE [van der Maaten and Hinton, 2008]. Following the idea of Booth et al. [Booth et al., 2016], we train a simple principal component analysis (PCA) model of the neutral faces in FaceWarehouse and EPFL3DFace, project the training samples onto that d-dimensional subspace and use t-SNE to generate a 2D visualization of that subspace. We then label the samples according to which database they belong to. Figure 6.9 shows the resulting visualization.

More specifically, we represent each shape as a vector $S = (x_0, y_0, z_0, ..., x_k, y_k, z_k)$, with $k =$ 5956, the number of vertices lying on the face, as defined in section 6.3.1. We then compute a PCA decomposition of the matrix whose rows are the shape vectors. Only the first 96 eigenvectors, which together explain more than 99% of the variance of the data, are kept. Each shape is then projected on the PCA basis, thus effectively reducing the original high number of dimensions of these. The new parametrization of the shapes in the PCA basis is the input to the t-SNE algorithm. t-SNE then projects the data to a low-dimensional subspace,

Figure 6.7 – Alignment results on two subjects and seven different facial expressions. (a) Color target (b) Rigidly aligned source (c) Result of the nonrigid alignment (d) Target without color (for better comparison).



(a)



(b)

Figure 6.8 – Evolution of the objective function and corresponding shapes during the optimization (a) Evolution of the individual terms in the objective function as well as the total cost for each iteration of the optimization. (b) The first row shows the distance to the target, normalized over the whole sequence and the second row the amplitude of the deformation for different steps of the optimization process, normalized in the same way.

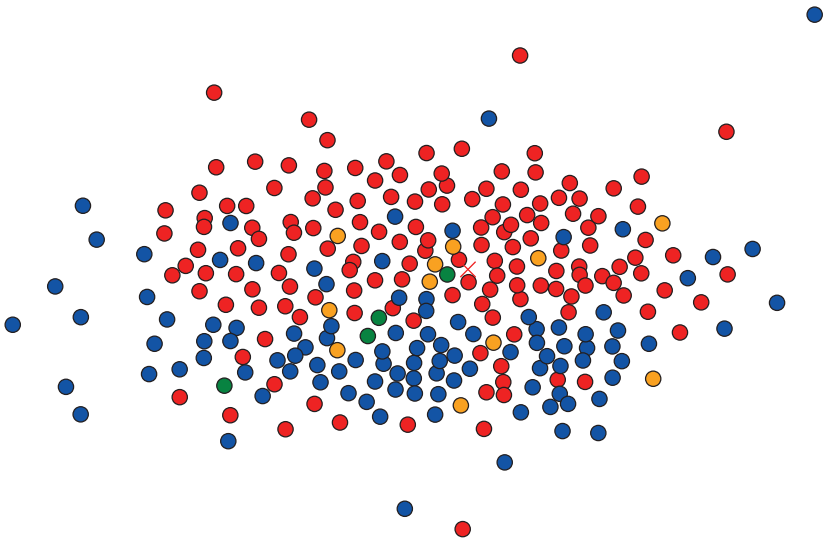


Figure 6.9 – Database subspaces visualization. ● FaceWarehouse, × Mean shape FaceWarehouse, ● Caucasian subjects from *EPFL3dFace*, ● Asian subjects form *EPFL3dFace*, ● South American subjects from *EPFL3dFace*.

typically 2D, while preserving similarities between data points and allows to visualize the structure of the data [van der Maaten and Hinton, 2008]. In this 2D space, we then label each point, which corresponds to each shape, according to the database that shape belongs to. For EPFL3DFace, we also chose to label the different ethnicities differently. This is not possible for FaceWarehouse, as we do not have the ground truth labels for the ethnicity of the subjects.

In the visualization in figure 6.9, shapes from different databases appear to be clustered and these clusters span different part of the subspace. Moreover, all the shapes from EPFL3DFace are obtained by nonrigidly deforming the mean shape of the corresponding expression in FaceWarehouse, represented as a cross in figure 6.9. This mean shape, the neutral expression in that case, is thus effectively deformed in a way that is complementary to the existing shapes in FaceWarehouse.

6.5 Conclusion

In this chapter, we introduce a new method to nonrigidly register a template to 3D surfaces. We take advantage of spectral geometry processing methods and propose to use Manifold Harmonics Transform (MHT) to constrain the deformation of the template, while enforcing

smoothness and reducing the number of parameters in the deformation model. More advanced use of the spectral nature of that deformation model needs to be further investigated. For example, it could be beneficial to select a different frequency band in which to deform the template, depending on the template, the level of details and the application. In our case, we show qualitatively that we obtain a reasonable level of details using only the first 500 spectral bases.

In addition, we propose to use an implicit surface representation based on multilevel partition of unity (MPU) for the target. This presents two main advantages: first, this new representation of the target surface allows to denoise the surface by approximating rather than interpolating it and fill in missing data. Second, the evaluation of the distance to the target is considerably simplified and is reduced to evaluating the implicit function, avoiding the need for correspondences.

Finally, we apply the proposed method on 3D facial scans in order to align them, or put them in dense correspondence. This is required to perform statistical analysis on the set of shapes and ultimately train a 3D statistical shape model. The nonrigidly registered set of shapes constitutes a new database of 3D facial expressions, EPFL3DFace, containing 120 subjects performing 35 different facial expressions and movements. This database is available to the research community upon request. We show that EPFL3DFace is complementary to the existing FaceWarehouse database and that both of them can be combined such that the number of subjects is increased by 80% and that they span a larger subspace.

EPFL3DFace and the presented spectral nonrigid alignment method constitute the first steps towards a 3D model of the face including a representative population and a large number of facial expressions and movements. From there, a statistical model can be learned using one of the many formulations that have been proposed, from the initial morphable model [Blanz and Vetter, 1999], based on a single dimension principal component analysis (PCA), to more recent multilinear models [Vlasic et al., 2005], or blendshape models [Weise et al., 2011]. Such a 3D face model can then be exploited in many different facial image analysis applications, such as the difficult tracheal intubation problem presented in part I of this thesis. More specifically, it acts as a prior to constrain the 3D reconstruction of the shape of the face from one or several 2D images. The morphological features would then be computed directly on the reconstructed 3D shape. This presents the advantages that these features can be invariant to the pose, as the reconstructed 3D shape is pose independent, and that several views can be combined in order to refine the 3D reconstruction.

Conclusions

In this final chapter, we review our contributions and main findings and discuss their benefits and limitations. We also present an outlook for future research perspectives, both in terms of the methodology and the medical application we targeted in this thesis, the automatic prediction of difficult tracheal intubation.

Summary and discussion of findings

In chapter 1, we have reviewed core components of a two dimensional (2D) facial image analysis pipeline, namely face detection and facial landmarks localization. We have provided a comprehensive introduction to the topic, by reviewing two face detectors, amongst which the Viola-Jones face detector, probably the most used real-time face detection algorithm in the past 15 years. We have provided a categorization of facial landmark localization methods and a review of four of them, amongst which the supervised descent method (SDM), a state-of-the-art regression-based method. These four methods were then benchmarked on two publicly available databases. In order to give a comprehension of the advantages and limitations of each method, we have discussed the results of the benchmark.

In chapter 3, we have presented a method to classify views of the mouth opening according to the visibility of the oropharyngeal structures, as defined in the modified Mallampati score. To the best of our knowledge, that is the first work proposing an automatic system for this task. Using features from a linear texture model, our method correctly classifies 95 out of 100 samples, and reaches 100% recall and precision for Mallampati class 4, which is an indicator of difficult intubation. The main limitation of the work, as presented in chapter 3, is that we assume that the facial landmarks are known. We have later demonstrated, in chapter 4, that these can indeed be localized and, thus, that automatic modified Mallampati score prediction can be integrated in a more complete system to predict the difficulty of tracheal intubation.

In chapter 4, we have proposed a fully automatic, facial morphometry based method to predict the difficulty of patients' tracheal intubation. Our method was developed and tested on the largest database of patients related to endotracheal intubation and reached an area under the curve (AUC) of 81% on a research-oriented scenario and 77.9% on a real-world scenario. We have demonstrated that the proposed method performs as well as state-of-the-

art multifactorial tests performed manually by experienced anesthesiologists. Yet, it does not require any measurement on the patient other than frontal and profile photographs, making it practical even for untrained personnel. At its current level of performance, this method already has the potential to reduce the costs, and increase the availability of such predictions, by not relying on qualified anesthesiologists with years of medical training. Nevertheless, we believe that, in order to reach its full potential and truly improve the patients' safety, an automatic method needs to reach superior performance. The main limitation for reaching such performance is the lack of data, including the images of the patients' face, but also a reliable ground-truth. This implies two things; first, more patients need to be recorded, and second, we believe that more work needs to be done in order to reach a reliable and objective definition of a difficult tracheal intubation, as discussed in chapter 2. The class imbalance between easy and difficult patients accentuates the need for more data. Indeed, even though we developed and validated the proposed method on more than 900 patients, only 60 of them were difficult to intubate. This shows that a large number of patients need to be recorded in order to get a sufficient number of difficult patients. In the next section, we discuss a few approaches to ease the data collection and increase the number of patients.

Another limitation of the methods presented in part I is the 2D nature of these methods. The head-pose variations in between patients can affect the normalization of the features and introduce noise in these. Moreover, the features from different views, typically from frontal and profile views of the face, are extracted independently, which might be suboptimal. These limitations have been the main motivation for the second part of this thesis and the development of a three dimensional (3D) model of the face.

For this purpose, in chapter 6, we have proposed a 3D nonrigid registration method based on spectral analysis on the mesh manifold. By constraining the deformation of a template in the frequency domain, we have shown that smoothness of the deformation can be ensured by optimizing it only in the low frequencies. This also allows to considerably reduce the number of free parameters in the optimization. In our experiment, this number was reduced by a factor of 92. We also proposed to represent the target mesh using an implicit surface representation, based on multilevel partition of unity (MPU). By using a local approximant, as opposed to non-local kernels in radial basis functions, in combination with an adaptive data structure, it is possible to efficiently represent surfaces for any given level of accuracy. As the focus of this work is the development of a 3D face model, adapted to the difficult tracheal intubation prediction, we have demonstrated the applicability of our nonrigid registration method on a new 3D facial expressions database, EPFL3DFace. This database contains 120 subjects, performing 35 different facial expressions and movements, recorded with a Microsoft Kinect®. One limitation of the methodology is the lack of comparison with other methods, for the nonrigid alignment method as well as for the implicit surface representation. The comparison is, indeed, difficult, as no ground truth is available in terms of alignment.

Future perspectives

In this section, we list a few perspectives for future research, which were identified in the scope of this thesis. First, we discuss a few research directions from a technical point-of-view to further improve our spectral nonrigid registration method and any resulting 3D model. Then, from a broader perspective, we suggest a few ideas to continue our work on the automatic prediction of difficult tracheal intubation.

On spectral nonrigid registration and 3D face models

From the methodological point-of-view, there is room for more in depth exploration of the influence of the different parameters of the method. As the method uses a spectral embedding, the choice of the frequencies, or the frequency bands, is completely open. In this work, we selected a low frequency band in order to ensure smoothness of the deformation, but different frequencies could be used with different properties. Perhaps some of the most important frequencies of a given input mesh could even be learned, which would allow to reduce further the number of parameters.

We also mentioned that our database of 120 subjects could, in principle, be augmented with nonrigidly registered scans from other databases. We believe that this would be an interesting research direction and could lead to a more complete model, or to different models that could adapt to different populations.

In the work presented in this thesis, we tackle the problem of nonrigid alignment, but we do not train a complete model with the aligned scans. Different formulations of statistical models have been proposed, from the initial morphable model [Blanz and Vetter, 1999], based on a single dimension principal component analysis (PCA), to more recent multilinear models [Vlasic et al., 2005], or blendshape models [Weise et al., 2011]. A thorough comparison of the expressive power of different statistical models would be of high value.

On automatic prediction of difficult tracheal intubation

From the application point-of-view, we saw that the lack of data, in terms of patients' recordings and ground-truth, was severely limiting the performance. Any development to ease the data collection and increase the number of patients would thus be very beneficial. A portable solution, for example on mobile, would allow for large-scale deployment of the data collection process across hospitals, in a multi-centric fashion. Technically, such a solution is not a big challenge and the obstacles would probably be more on a regulatory level.

Finally, a direct extension would be the application of the new 3D facial model to the difficult tracheal intubation problem. As discussed extensively already, extracting the morphological features directly from a 3D representation of the face presents significant advantages. Moreover, the applications of such a 3D model of the face are not limited to a specific problem in

anesthesiology. Many current applications of facial image analysis can benefit from the availability of such a model. As an example, recent work tries to uncover the relationships between facial variation, modeled with a linear statistical model of the face, and a subset of candidate genes [Claes et al., 2014]. This approach aims at identifying genes affecting normal-range facial features as well as predicting the appearance of a face from genetic markers. Having both the ability to reconstruct the 3D shapes of over 2700 patients' faces, recorded in the first part of this thesis, and the access to genomic information of these same patients, through the hospital biobank, potentially enables contributions in this (futuristic) research topic.

Bibliography

[Adnet et al., 1997] Adnet, F., Borron, S. W., Racine, S. X., Clemessy, J.-L., Fournier, J.-L., Plaisance, P. and Lapandry, C. (1997). The Intubation Difficulty Scale (IDS). Anesthesiology *87*, 1290–1297.

[Ahn and Picard, 2014] Ahn, H. I. and Picard, R. W. (2014). Measuring affective-cognitive experience and predicting market success. IEEE Transactions on Affective Computing *5*, 173–186.

[Alberto et al., 2012] Alberto, K., Mora, F. and Odobez, J.-M. (2012). Gaze Estimation from Multimodal Kinect Data. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4321–4326,.

[Allen et al., 2003] Allen, B., Curless, B. and Popović, Z. (2003). The space of human body shapes. ACM Transactions on Graphics *22*, 587.

[Amberg et al., 2009] Amberg, B., Blake, A. and Vetter, T. (2009). On compositional Image Alignment, with an application to Active Appearance Models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1714–1721,.

[Amberg et al., 2007] Amberg, B., Romdhani, S. and Vetter, T. (2007). Optimal Step Nonrigid ICP Algorithms for Surface Registration. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–8, IEEE.

[Anderson et al., 2013] Anderson, R., Stenger, B., Wan, V. and Cipolla, R. (2013). Expressive visual text-to-speech using active appearance models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3382–3389,.

[Apfelbaum et al., 2013] Apfelbaum, J. L., Hagberg, C. A., Caplan, R. A., Blitt, C. D., Connis, R. T. and Nickinovich, D. G. (2013). Practice Guidelines for Management of the Difficult Airway: An Updated Report. Technical Report 2.

[Arar et al., 2012a] Arar, N. M., Bekmezci, N. K., Fatma, G., Bilgisayar, M. and Ekenel, H. K. (2012a). Real-time Face Swapping in Video Sequences: Magic Mirror. In Proceedings of the 7th International Conference on Computer Vision Theory and Applications.

[Arar et al., 2012b] Arar, N. M., Gao, H., Ekenel, H. K. and Akarun, L. (2012b). Selection and combination of local Gabor classifiers for robust face verification. In 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS) pp. 297–302, IEEE.

[Arar et al., 2015] Arar, N. M., Hua Gao and Thiran, J.-P. (2015). Robust gaze estimation based on adaptive fusion of multiple cameras. In Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) pp. 1–7, IEEE.

[Arné et al., 1998] Arné, J., Descoins, P., Fusciardi, J., Ingrand, P., Ferrier, B., Boudigues, D. and Ariès, J. (1998). Preoperative assessment for difficult intubation in general and ENT surgery: predictive value of a clinical multivariate risk index. British journal of anaesthesia *80*, 140–6.

[Arun et al., 1987] Arun, K. S., Huang, T. S. and Blostein, S. D. (1987). Least-Squares Fitting of Two 3-D Point Sets. IEEE Transactions on Pattern Analysis and Machine Intelligence *9*, 698–700.

[Ashraf et al., 2010] Ashraf, A. B., Lucey, S. and Chen, T. (2010). Fast image alignment in the Fourier domain. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2480–2487, IEEE.

[Asteriadis et al., 2009] Asteriadis, S., Nikolaidis, N. and Pitas, I. (2009). Facial feature detection using distance vector fields. Pattern Recognition *42*, 1388–1398.

[Asthana et al., 2009] Asthana, A., Goecke, R., Quadrianto, N. and Gedeon, T. (2009). Learning based automatic face annotation for arbitrary poses and expressions from frontal images only. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1635–1642,.

[Asthana et al., 2011] Asthana, A., Lucey, S. and Goecke, R. (2011). Regression based automatic face annotation for deformable model building. Pattern Recognition *44*, 2598–2613.

[Asthana et al., 2013] Asthana, A., Zafeiriou, S., Cheng, S. and Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3444–3451,.

[Aziz et al., 2011] Aziz, M. F., Healy, D., Kheterpal, S., Fu, R. F., Dillman, D. and Brambrink, A. M. (2011). Routine clinical practice effectiveness of the Glidescope in difficult airway management: an analysis of 2,004 Glidescope intubations, complications, and failures from two institutions. Anesthesiology *114*, 34–41.

[Baker et al., 2009] Baker, P. A., Depuydt, A. and Thompson, J. M. D. (2009). Thyromental distance measurement - fingers don't rule. Anaesthesia *64*, 878–82.

[Baker and Matthews, 2004] Baker, S. and Matthews, I. (2004). Lucas-Kanade 20 Years On: A Unifying Framework. International Journal of Computer Vision *56*, 221–255.

[Baker et al., 2004] Baker, S., Matthews, I. and Schneider, J. (2004). Automatic construction of active appearance models as an image coding problem. IEEE Transactions on Pattern Analysis and Machine Intelligence *26*, 1380–4.

[Baltrusaitis et al., 2012] Baltrusaitis, T., Robinson, P. and Morency, L.-P. (2012). 3D Constrained Local Model for rigid and non-rigid facial tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2610–2617,.

[Batur and Hayes, 2003] Batur, A. U. and Hayes, M. H. (2003). A novel convergence scheme for active appearance models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 359–366,.

[Batur and Hayes, 2005] Batur, A. U. and Hayes, M. H. (2005). Adaptive active appearance models. IEEE Transactions on Image Processing *14*, 1707–21.

[Baynam et al., 2011] Baynam, G., Claes, P., Craig, J. M., Goldblatt, J., Kung, S., Le Souef, P. and Walters, M. (2011). Intersections of epigenetics, twinning and developmental asymmetries: insights into monogenic and complex diseases and a role for 3D facial analysis. Twin Research and Human Genetics *14*, 305–315.

[Belhumeur et al., 2011] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J. and Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 545–552, IEEE.

[Belhumeur et al., 2013] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J. and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence *35*, 2930–2940.

[Besl and McKay, 1992] Besl, P. and McKay, N. D. (1992). A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence *14*, 239–256.

[Beumier and Acheroy, 2001] Beumier, C. and Acheroy, M. (2001). Face verification from 3D and grey level clues. Pattern Recognition Letters *22*, 1321–1329.

[Black and Jepson, 1998] Black, M. J. and Jepson, A. D. (1998). EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. International Journal of Computer Vision *26*, 63–84.

[Blais and Levine, 1995] Blais, G. and Levine, M. (1995). Registering Multiview Range Data to Create 3D Computer Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence *17*, 820–824.

[Blanz et al., 2004] Blanz, V., Mehl, A., Vetter, T. and Seidel, H.-P. (2004). A statistical method for robust 3D surface reconstruction from sparse data. In Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. pp. 293–300, IEEE.

[Blanz and Vetter, 1999] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In Proceedings of the conference on Computer graphics and interactive techniques (SIGGRAPH) pp. 187–194, ACM Press, New York, USA.

[Bolkart and Wuhrer, 2015] Bolkart, T. and Wuhrer, S. (2015). A Groupwise Multilinear Correspondence Optimization for 3D Faces. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 3604–3612, IEEE.

[Booth et al., 2016] Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A. and Dunaway, D. (2016). A 3D Morphable Model learnt from 10'000 faces. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE.

[Botsch et al., 2010] Botsch, M., Kobbelt, L., Pauly, M., Alliez, P. and Lévy, B. (2010). Polygon Mesh Processing, vol. 1,. A K Peters, Ltd.

[Botsch and Sorkine, 2008] Botsch, M. and Sorkine, O. (2008). On Linear Variational Surface Deformation Methods. IEEE Transactions on Visualization and Computer Graphics *14*, 213–230.

[Bottino and Laurentini, 2010] Bottino, A. and Laurentini, A. (2010). The Analysis of Facial Beauty: An Emerging Area of Research in Pattern Analysis. In International Conference on Image Analysis and Recognition pp. 425–435.

[Bradski, 2000] Bradski, G. (2000). The opencv library. Doctor Dobbs Journal *25*, 120–126.

[Breiman, 2001] Breiman, L. (2001). Random Forests. Machine Learning *45*, 5–32.

[Breitenstein et al., 2008] Breitenstein, M. D., Kuettel, D., Weise, T., van Gool, L. and Pfister, H. (2008). Real-time face pose estimation from single range images. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) number 1 pp. 1–8, IEEE.

[Bronstein et al., 2006] Bronstein, A. M., Bronstein, M. M. and Kimmel, R. (2006). Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. Proceedings of the National Academy of Sciences *103*, 1168–1172.

[Bronstein et al., 2007a] Bronstein, A. M., Bronstein, M. M. and Kimmel, R. (2007a). Calculus of nonrigid surfaces for geometry and texture manipulation. IEEE Transactions on Visualization and Computer Graphics *13*, 902–913.

[Bronstein et al., 2007b] Bronstein, A. M., Bronstein, M. M. and Kimmel, R. (2007b). Expression-Invariant Representations of Faces. IEEE Transactions on Image Processing *16*, 188–197.

[Bronstein et al., 2008] Bronstein, A. M., Bronstein, M. M. and Kimmel, R. (2008). Numerical Geometry of Non-Rigid Shapes. Monographs in Computer Science, springer edition, Springer New York, New York, NY.

[Brunton et al., 2014] Brunton, A., Salazar, A., Bolkart, T. and Wuhrer, S. (2014). Review of statistical shape spaces for 3D data with comparative analysis for human faces. Computer Vision and Image Understanding *128*, 1–17.

[Burgos-Artizzu et al., 2013] Burgos-Artizzu, X. P., Perona, P. and Dollar, P. (2013). Robust face landmark estimation under occlusion. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 1513–1520,.

[Cao et al., 2013] Cao, C., Weng, Y., Lin, S. and Zhou, K. (2013). 3D Shape Regression for Real-time Facial Animation. ACM Transactions on Graphics *32*, 41.1–41.10.

[Cao et al., 2012] Cao, X., Wei, Y., Wen, F. and Sun, J. (2012). Face alignment by Explicit Shape Regression. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2887–2894, Ieee.

[Caplan et al., 2003] Caplan, R., Benumof, J. and Berry, F. (2003). Practice Guidelines for Management of the Difficult. Anesthesiology *98*, 1269–1277.

[Cattano et al., 2013] Cattano, D., Killoran, P. V., Iannucci, D., Maddukuri, V., Altamirano, A. V., Sridhar, S., Seitan, C., Chen, Z. and Hagberg, C. A. (2013). Anticipation of the difficult airway: preoperative airway assessment, an educational and quality improvement tool. British journal of anaesthesia *111*, 1–10.

[Cattano et al., 2004] Cattano, D., Panicucci, E., Paolicchi, a., Forfori, F., Giunta, F. and Hagberg, C. (2004). Risk factors assessment of the difficult airway: an italian survey of 1956 patients. Anesthesia and analgesia *99*, 1774–9, table of contents.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology *2*, 27:1–27:27.

[Chang et al., 2010] Chang, W., Huang, Q.-X., Li, H., Mitra, N. J., Pauly, M. and Wand, M. (2010). Geometric Registration for Deformable Shapes. In Eurographics (Tutorials).

[Chen Cao et al., 2014] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong and Kun Zhou (2014). FaceWarehouse: A 3D Facial Expression Database for Visual Computing. IEEE Transactions on Visualization and Computer Graphics *20*, 413–425.

[Chew et al., 2011] Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F. and Sridharan, S. (2011). Person-independent facial expression detection using Constrained Local Models. In Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (FG) pp. 915–920,.

[Claes et al., 2012] Claes, P., Daniels, K., Walters, M., Clement, J., Vandermeulen, D. and Suetens, P. (2012). Dysmorphometrics: The modelling of morphological abnormalities. Theoretical biology & medical modelling *9*, 5.

[Claes et al., 2014] Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., McEvoy, B., Bauchet, M., Zaidi, A. a., Yao, W., Tang, H., Barsh, G. S., Absher, D. M., Puts, D. a., Rocha, J., Beleza, S., Pereira, R. W., Baynam, G., Suetens, P., Vandermeulen, D., Wagner, J. K., Boster, J. S. and Shriver, M. D. (2014). Modeling 3D Facial Shape from DNA. PLoS Genetics *10*, e1004224.

[Colombo et al., 2011] Colombo, A., Cusano, C. and Schettini, R. (2011). UMB-DB: A database of partially occluded 3D faces. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) pp. 2113–2119, IEEE.

[Connor and Segal, 2011] Connor, C. W. and Segal, S. (2011). Accurate classification of difficult intubation by computerized facial analysis. Anesthesia and analgesia *112*, 84–93.

[Cook and Macdougall-Davis, 2012] Cook, T. M. and Macdougall-Davis, S. R. (2012). Complications and failure of airway management. British journal of anaesthesia *109 Suppl*, i68–i85.

[Cootes et al., 2012] Cootes, T., Ionita, M., Lindner, C. and Sauer, P. (2012). Robust and accurate shape model fitting using random forest regression voting. In Proceedings of European Conference on Computer Vision (ECCV) pp. 1–14,.

[Cootes et al., 1992] Cootes, T. F., Cooper, D., Taylor, C. and Graham, J. (1992). Trainable method of parametric shape description. Image and Vision Computing *10*, 289–294.

[Cootes et al., 1998] Cootes, T. F., Edwards, G. and Taylor, C. J. (1998). A Comparative Evaluation of Active Appearance Model Algorithms. In Procedings of the British Machine Vision Conference (BMVC) pp. 68.1–68.10,.

[Cootes et al., 2001] Cootes, T. F., Edwards, G. J. and Taylor, C. (2001). Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence *23*, 681–685.

[Cootes et al., 1998] Cootes, T. F., Edwards, G. J. and Taylor, C. J. (1998). Active appearance models. In Proceedings of European Conference on Computer Vision (ECCV) pp. 484–498,.

[Cootes and Taylor, 2004] Cootes, T. F. and Taylor, C. (2004). Statistical Models of Appearance for Computer Vision. Technical report University of Manchester.

[Cootes and Taylor, 1993] Cootes, T. F. and Taylor, C. J. (1993). Active Shape Model Search using Local Grey-Level Models : A Quantitative Evaluation. In Proceedings of the British Machine Vision Conference (BMVC) pp. 639–648,.

[Cootes and Taylor, 2001] Cootes, T. F. and Taylor, C. J. (2001). Constrained Active Appearance Models. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 748–754,.

[Cootes and Taylor, 2006] Cootes, T. F. and Taylor, C. J. (2006). An Algorithm for Tuning an Active Appearance Model to New Data. In Procedings of the British Machine Vision Conference (BMVC) pp. 94.1–94.10,.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. and Graham, J. (1995). Active Shape Models - Their Training and Application. Computer Vision and Image Understanding *61*, 38–59.

[Cootes et al., 1994] Cootes, T. F., Taylor, C. J. and Lanitis, A. (1994). Active shape models: Evaluation of a multi-resolution method for improving image search. In Proceedings of the British Machine Vision Conference (pp. 327–338,.

[Cootes et al., 2002] Cootes, T. F., Wheeler, G. V., Walker, K. N. and Taylor, C. J. (2002). View-based active appearance models. Image and Vision Computing *20*, 657–664.

[Cormack and Lehane, 1984] Cormack, R. S. and Lehane, J. R. (1984). Difficult tracheal intubation in obstetrics. Anaesthesia *39*, 1105–1111.

[Coughlan and Ferreira, 2002] Coughlan, J. M. and Ferreira, S. J. (2002). Finding deformable shapes using Loopy belief propagation. In Proceedings of European Conference on Computer Vision (ECCV) pp. 453–468,.

[Cristinacce and Cootes, 2006] Cristinacce, D. and Cootes, T. (2006). Facial Feature Detection and Tracking with Automatic Template Selection. In Proceedings of International Conference on Automatic Face and Gesture Recognition (FG) pp. 429–434, IEEE.

[Cristinacce et al., 2004] Cristinacce, D., Cootes, T. and Scott, I. (2004). A Multi-Stage Approach to Facial Feature Detection. In Procedings of the British Machine Vision Conference (BMVC) pp. 30.1–30.10,.

[Cristinacce and Cootes, 2003] Cristinacce, D. and Cootes, T. F. (2003). Facial feature detection using AdaBoost with shape constraints. In Procedings of the British Machine Vision Conference (BMVC) pp. 24.1–24.10,.

[Cristinacce and Cootes, 2004] Cristinacce, D. and Cootes, T. F. (2004). A comparison of shape constrained facial feature detectors. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG) pp. 375–380,.

[Cristinacce and Cootes, 2006] Cristinacce, D. and Cootes, T. F. (2006). Feature Detection and Tracking with Constrained Local Models. In Procedings of the British Machine Vision Conference (BMVC) pp. 95.1–95.10, British Machine Vision Association.

[Cristinacce and Cootes, 2007] Cristinacce, D. and Cootes, T. F. (2007). Boosted Regression Active Shape Models. In Procedings of the British Machine Vision Conference (BMVC) pp. 79.1–79.10, British Machine Vision Association.

[Cristinacce and Cootes, 2008] Cristinacce, D. and Cootes, T. F. (2008). Automatic feature localisation with constrained local models. Pattern Recognition *41*, 3054–3067.

[Cuendet et al., 2015] Cuendet, G., Schoettker, P., Yuce, A., Sorci, M., Gao, H., Perruchoud, C. and Thiran, J.-P. (2015). Facial Image Analysis for Fully-Automatic Prediction of Difficult Endotracheal Intubation. IEEE Transactions on Biomedical Engineering *63*, 328–339.

[Cuendet et al., 2017] Cuendet, G. L., Ecabert, C., Zimmermann, M., Ekenel, H. K., Thiran, J.-p. and Member, S. (2017). 3D Spectral Nonrigid Registration of Facial Expression Scans. IEEE Transactions on Visualization and Computer Graphics *submitted*, 1–13.

[Cuendet et al., 2012] Cuendet, G. L., Yüce, A., Sorci, M., Schoettker, P., Perruchoud, C. and Thiran, J.-P. (2012). Automatic Mallampati Classification Using Active Appearance Models. In Proc. of International Workshop on Pattern Recognition for Healthcare Analytics.

[Curless and Levoy, 1996] Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96 pp. 303–312, ACM Press, New York, New York, USA.

[Dalal and Triggs, 2010] Dalal, N. and Triggs, B. (2010). Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) vol. 1, pp. 886–893, IEEE.

[Dantone et al., 2012] Dantone, M., Gall, J., Fanelli, G. and Van Gool, L. (2012). Real-time facial feature detection using conditional regression forests. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2578–2585, Ieee.

[Davis, 1975] Davis, L. S. (1975). A survey of edge detection techniques. Computer Graphics and Image Processing *4*, 248–270.

[Dedeoglu et al., 2006] Dedeoglu, G., Baker, S. and Kanade, T. (2006). Resolution-aware fitting of Active Appearance Models to low resolution images. In Proceedings of European Conference on Computer Vision (ECCV) pp. 83–97,.

[Dev, 1974] Dev, P. (1974). Segmentation processes in visual perception: A cooperative neural model. Technical report COINS Technical Report 74C-5, University of Massachusetts at Amherst.

[Diemunsch et al., 2008] Diemunsch, P., Langeron, O., Richard, M. and Lenfant, F. (2008). Prédiction et définition de la ventilation au masque difficile et de l'intubation difficile. Annales françaises d'anesthésie et de rénimation *27*, 3–14.

[Ding et al., 2016] Ding, C., Tao, D., Systems, I. and Technology, I. (2016). A Comprehensive Survey on Pose-Invariant Face Recognition. ACM Transactions on Intelligent Systems and Technology *7*.

[Ding and Martinez, 2008] Ding, L. and Martinez, A. M. (2008). Precise detailed detection of faces and facial features. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Ding and Martinez, 2010] Ding, L. and Martinez, A. M. (2010). Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. IEEE Transactions on Pattern Analysis and Machine Intelligence *32*, 2022–2038.

[Dong et al., 2009] Dong, Y., Hu, Z., Uchimura, K. and Murayama, N. (2009). Driver inattention monitoring system for intelligent vehicles: A review. In 2009 IEEE Intelligent Vehicles Symposium vol. 8555, pp. 875–880, IEEE.

[Donner et al., 2006] Donner, R., Reiter, M., Langs, G., Peloschek, P. and Bischof, H. (2006). Fast active appearance model search using canonical correlation analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence *28*, 1690–1694.

[Eberhart et al., 2005] Eberhart, L. H. J., Arndt, C., Cierpka, T., Schwanekamp, J., Wulf, H. and Putzke, C. (2005). The reliability and validity of the upper lip bite test compared with the Mallampati classification to predict difficult laryngoscopy: an external prospective evaluation. Anesthesia and analgesia *101*, 284–289.

[Elkan, 2001] Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01) pp. 973–978,.

[Faltemier et al., 2007] Faltemier, T. C., Bowyer, K. W. and Flynn, P. J. (2007). Using a multi-instance enrollment representation to improve 3D face recognition. In IEEE Conference on Biometrics: Theory, Applications and Systems, BTAS'07.

[Fanelli et al., 2012] Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L. and Gool, L. (2012). Random Forests for Real Time 3D Face Analysis. International Journal of Computer Vision *101*, 437–458.

[Fanelli et al., 2013] Fanelli, G., Dantone, M. and Gool, L. V. (2013). Real Time 3D Face Alignment with Random Forests-based Active Appearance Models. In Proceedings of International Conference on Automatic Face and Gesture Recognition (FG) pp. 1–8,.

[Fanelli et al., 2010] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T. and Van Gool, L. (2010). A 3-D Audio-Visual Corpus of Affective Communication. IEEE Transactions on Multimedia *12*, 591–598.

[Fanelli et al., 2012] Fanelli, G., Gall, J. and Van Gool, L. (2012). Real Time 3D Head Pose Estimation: Recent Achievements and Future Challenges. In Proceedings of the International Symposium on Communications, Control and Signal Processing (ISCCSP), Rome.

[Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence *32*, 1627–45.

[Felzenszwalb and Huttenlocher, 2005] Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial Structures for Object Recognition. International Journal of Computer Vision *61*, 55–79.

[Fischler and Elschlager, 1973] Fischler, M. A. and Elschlager, R. A. (1973). The Representation and Matching of Pictorial Structures Representation. IEEE Transactions on Computers *C-22*, 67–92.

[Fleuret and Geman, 2001] Fleuret, F. and Geman, D. (2001). Coarse-to-Fine Face Detection. International Journal of Computer Vision *41*, 85–107.

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. Computational learning theory *55*, 119–139.

[Fritscherova et al., 2011] Fritscherova, S., Adamus, M., Dostalova, K., Koutna, J., Hrabalek, L., Zapletalova, J., Uvizl, R. and Janout, V. (2011). Can difficult intubation be easily and rapidly predicted? Biomedical Papers *155*, 165–172.

[Fu et al., 2010] Fu, Y., Guo, G. and Huang, T. S. (2010). Age synthesis and estimation via faces: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence *32*, 1955–76.

[Galar et al., 2012] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) *42*, 463–484.

[Gao et al., 2014] Gao, H., Yuce, A. and Thiran, J.-P. (2014). Detecting Emotional Stress from Facial Expressions for Driving Safety. In Proc. of International Conference on Image Processing (ICIP) vol. 1,.

[Gonzalez-Mora et al., 2007] Gonzalez-Mora, J., De la Torre, F., Murthi, R., Guil, N. and Zapata, E. L. (2007). Bilinear Active Appearance Models. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 1–8,.

[Goodall, 1991] Goodall, C. (1991). Procrustes Methods in the Statistical Analysis of Shape. Journal of the Royal Statistical Society *53*, 285–339.

[Green, 1984] Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation , and some Robust and Resistant Alternatives. Journal of the Royal Statistical Society. Series B (Methodological) *46*, 149–192.

[Gross et al., 2005] Gross, R., Matthews, I. and Baker, S. (2005). Generic vs. person specific active appearance models. Image and Vision Computing *23*, 1080–1093.

[Gross et al., 2010] Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S. (2010). Multi-PIE. Image and Vision Computing *28*, 807–813.

[Gu and Kanade, 2008] Gu, L. and Kanade, T. (2008). A Generative Shape Regularization Model for Robust Face Alignment. In Proceedings of IEEE European Conference on Computer Vision (ECCV) pp. 413–426,.

[Gupta et al., 2010] Gupta, S., Castleman, K. R., Markey, M. K. and Bovik, A. C. (2010). Texas 3D Face Recognition Database. In 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI) pp. 97–100, IEEE.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. N. (2002). Gene selection for cancer classification using support vector machines. Machine Learning *46*, 389–422.

[Hamsici and Martinez, 2009] Hamsici, O. C. and Martinez, A. M. (2009). Active Appearance Models with Rotation Invariant Kernels. In Proceedings of IEEE 12th International Conference on Computer Vision (ICCV) pp. 1003–1009, IEEE.

[Hansen et al., 2011] Hansen, M. F., Fagertun, J. and Larsen, R. (2011). Elastic appearance models. In Proceedings of British Machine Vision Conference (BMVC) pp. 1–12,.

[He and Garcia, 2009] He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering *21*, 1263–1284.

[Heard et al., 2009] Heard, A. M. B., Green, R. J., Eakins, P., Heard, M. B., Green, R. J. and Eakins, P. (2009). The formulation and introduction of a 'can't intubate, can't ventilate' algorithm into clinical practice. Anaesthesia *64*, 601–608.

[Heseltine et al., 2008] Heseltine, T., Pears, N. and Austin, J. (2008). Three-dimensional face recognition using combinations of surface feature map subspace components. Image and Vision Computing *26*, 382–396.

[Horn and Schunk, 1981] Horn, B. K. P. and Schunk, B. G. (1981). Determining Optical Flow. Artificial Intelligence *17*, 185–203.

[Hou et al., 2001] Hou, X., Li, S., Zhang, H. and Cheng, Q. (2001). Direct appearance models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 828–833,.

[Hove et al., 2007] Hove, L. D., Steinmetz, J., Christoffersen, J. K., Møller, A., Nielsen, J. and Schmidt, H. (2007). Analysis of deaths related to anesthesia in the period 1996-2004 from closed claims registered by the Danish Patient Insurance Association. Anesthesiology *106*, 675–80.

[Hsu and Lin, 2002] Hsu, C.-W. and Lin, C.-J. (2002). Errata to "A comparison of methods for multiclass support vector machines". IEEE transactions on neural networks *13*, 415–425.

[Huang et al., 2012] Huang, C., Ding, X. and Fang, C. (2012). Pose robust face tracking by combining view-based AAMs and temporal filters. Computer Vision and Image Understanding *116*, 777–792.

[Huang et al., 2006] Huang, J., Shi, X., Liu, X., Zhou, K., Wei, L.-Y., Teng, S.-H., Bao, H., Guo, B. and Shum, H.-Y. (2006). Subspace gradient domain mesh deformation. ACM Transactions on Graphics *25*, 1126.

[Huang, 1981] Huang, T. S. (1981). Image sequence analysis. Springer Verlag, Berlin, Heidelberg.

[Huang et al., 2007] Huang, Y., Liu, Q. and Metaxas, D. (2007). A component based deformable model for generalized face alignment. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 0–7,.

[Huber et al., 2016] Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, W. P., Christmas, W. J., Rätsch, M. and Kittler, J. (2016). A Multiresolution 3D Morphable Face Model and Fitting Framework. In Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications pp. 79–86, SCITEPRESS - Science and and Technology Publications.

[Hung et al., 2016] Hung, O., Law, J. A., Morris, I. and Murphy, M. (2016). Airway Assessment Before Intervention. Anesthesia & Analgesia *122*, 1752–1754.

[Ichim et al., 2015] Ichim, A. E., Bouaziz, S. and Pauly, M. (2015). Dynamic 3D avatar creation from hand-held video input. ACM Transactions on Graphics *34*, 45:1–45:14.

[Izadi et al., 2011] Izadi, S., Davison, A., Fitzgibbon, A., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S. and Freeman, D. (2011). KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moveing Depth Camera. In Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11 pp. 559–568, ACM Press, New York, New York, USA.

[Jain, 1989] Jain, A. K. (1989). Fundamentals of digital image processing. Prentice-Hall, Inc.

[Jaiswal and Valstar, 2016] Jaiswal, S. and Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 1–8, IEEE.

[Juan et al., 2002] Juan, E. J., Mansfield, J. P. and Wodicka, G. R. (2002). Miniature acoustic guidance system for endotracheal tubes. IEEE Transactions on Biomedical Engineering *49*, 584–596.

[Kahraman et al., 2007] Kahraman, F., Gokmen, M., Darkner, S. and Larsen, R. (2007). An Active Illumination and Appearance (AIA) Model for Face Alignment. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–7, IEEE.

[Kakadiaris et al., 2007] Kakadiaris, I. A., Passalis, G., Toderici, G., Murtuza, M. N., Lu, Y., Karampatziakis, N. and Theoharis, T. (2007). Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. IEEE Transactions on Pattern Analysis and Machine Intelligence *29*, 640–649.

[Kalal et al., 2010] Kalal, Z., Mikolajczyk, K. and Matas, J. (2010). Face-TLD: Tracking-learning-detection applied to faces. In Proceedings of the International Conference on Image Processing, ICIP pp. 3789–3792,.

[Kazemi and Sullivan, 2011] Kazemi, V. and Sullivan, J. (2011). Face alignment with part-based modeling. In Proceedings of British Machine Vision Conference (BMVC) pp. 27.1–27.10,.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1867–1874, IEEE.

[Khan et al., 2003] Khan, Z. H., Kashfi, A. and Ebrahimkhani, E. (2003). A comparison of the upper lip bite test (a simple new technique) with modified Mallampati classification in predicting difficulty in endotracheal intubation: A prospective blinded study. Anesthesia and Analgesia *96*, 595–599.

[Khan et al., 2009] Khan, Z. H., Mohammadi, M., Rasouli, M. R., Farrokhnia, F. and Khan, R. H. (2009). The diagnostic value of the upper lip bite test combined with sternomental distance, thyromental distance, and interincisor distance for prediction of easy laryngoscopy and intubation: a prospective study. Anesthesia and analgesia *109*, 822–4.

[Kinoshita et al., 2012] Kinoshita, K., Konishi, Y., Kawade, M. and Murase, H. (2012). Facial model fitting based on perturbation learning and it's evaluation on challenging real-world diversities images. In Proceedings of European Conference on Computer Vision Workshop (ECCVW) pp. 153–162,.

[Kosilek et al., 2015] Kosilek, R. P., Frohner, R., Würtz, R. P., Berr, C. M., Schopohl, J., Reincke, M. and Schneider, H. J. (2015). Diagnostic use of facial image analysis software in endocrine and genetic disorders: Review, current results and future perspectives. European Journal of Endocrinology *173*, M39–M44.

[Kozakaya et al., 2008a] Kozakaya, T., Shibata, T., Takeguchi, T. and Nishiura, M. (2008a). Fully automatic feature localization for medical images using a global vector concentration approach. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Kozakaya et al., 2008b] Kozakaya, T., Shibata, T., Yuasa, M. and Yamaguchi, O. (2008b). Facial feature localization using weighted vector concentration approach. In Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (FG) pp. 1–6, IEEE.

[Kozakaya et al., 2010] Kozakaya, T., Shibata, T., Yuasa, M. and Yamaguchi, O. (2010). Facial feature localization using weighted vector concentration approach. Image and Vision Computing *28*, 772–780.

[Krage et al., 2010] Krage, R., van Rijn, C., Van Groeningen, D., Loer, S. A., Schwarte, L. A. and Schober, P. (2010). Cormack-Lehane classification revisited. British journal of anaesthesia *105*, 220–7.

[Krobbuaban et al., 2005] Krobbuaban, B., Diregpoke, S., Kumkeaw, S. and Tanomsat, M. (2005). The predictive value of the height ratio and thyromental distance: four predictive tests for difficult laryngoscopy. Anesthesia and Analgesia *101*, 1542–5.

[Kumar et al., 2009] Kumar, N., Berg, A. C., Belhumeur, P. N. and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In International Conference on Computer Vision (ICCV) pp. 365–372, IEEE.

[Langeron et al., 2012] Langeron, O., Cuvillon, P., Ibanez-Esteve, C., Lenfant, F., Riou, B. and Le Manach, Y. (2012). Prediction of Difficult Tracheal Intubation: Time for a Paradigm Change. Anesthesiology *117*, 1223–1233.

[Le et al., 2012] Le, V., Brandt, J., Lin, Z., Bourdev, L. and Huang, T. S. (2012). Interactive Facial Feature Localization. In Proceedings of European Conference on Computer Vision (ECCV) pp. 679–692,.

[Lee et al., 2006] Lee, A., Fan, L. T. Y., Gin, T., Karmakar, M. K., Kee, W. D. N. and Ngan Kee, W. D. (2006). A systematic review (meta-analysis) of the accuracy of the Mallampati tests to predict the difficult airway. Anesthesia and analgesia *102*, 1867–78.

[Lee and Kim, 2009] Lee, H. S. and Kim, D. (2009). Tensor-based AAM with continuous variation estimation: Application to variation-robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence *31*, 1102–1116.

[Lévy and Zhang, 2010] Lévy, B. and Zhang, H. (2010). Spectral Mesh Processing. In ACM SIGGRAPH Course Notes.

[Li et al., 2009] Li, H., Adams, B., Guibas, L. J. and Pauly, M. (2009). Robust single-view geometry and motion reconstruction. ACM Transactions on Graphics *28*, 1.

[Li et al., 2011] Li, Y., Gu, L. and Kanade, T. (2011). Robustly aligning a shape model and its application to car alignment of unknown pose. IEEE Transactions on Pattern Analysis and Machine Intelligence *33*, 1860–1876.

[Liang et al., 2006a] Liang, L., Wen, F., Tang, X. and Xu, Y.-q. (2006a). An Integrated Model for Accurate A Two-Level Shape Model. In Proceedings of European Conference on Computer Vision (ECCV) pp. 333–346,.

[Liang et al., 2006b] Liang, L., Wen, F., Xu, Y. Q., Tang, X. and Shum, H. Y. (2006b). Accurate face alignment using shape constrained Markov network. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) vol. 1, pp. 1313–1320,.

[Liang et al., 2008] Liang, L., Xiao, R., Wen, F. and Sun, J. (2008). Face Alignment Via Component-Based Discriminative Search. In Proceedings of European pp. 72–85, Springer Berlin Heidelberg, Berlin, Heidelberg.

[Liu, 2009] Liu, X. (2009). Discriminative face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence *31*, 1941–54.

[Liu et al., 2006] Liu, X., Tu, P. and Wheeler, F. W. (2006). Face Model Fitting on Low Resolution Images. In Procedings of the British Machine Vision Conference (BMVC) pp. 110.1–110.10,.

[López et al., 2013] López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences *250*, 113–141.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision *60*, 91–110.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. Proceedings of Imaging Understanding Workshop *130*, 121–130.

[Lucey et al., 2013] Lucey, S., Navarathna, R., Ashraf, A. B. and Sridharan, S. (2013). Fourier Lucas-Kanade algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence *35*, 1383–1396.

[Lucey et al., 2009] Lucey, S., Wang, Y., Cox, M., Sridharan, S. and Cohn, J. F. (2009). Efficient constrained local model fitting for non-rigid face alignment. Image and Vision Computing *27*, 1804–1813.

[Lundstrøm et al., 2011] Lundstrøm, L. H., Vester-Andersen, M., Møller, M., Charuluxananan, S., L'Hermite, J., Wetterslev, J., Møller, A. M., Charuluxananan, S., L'Hermite, J. and Wetterslev, J. (2011). Poor prognostic value of the modified Mallampati score: a meta-analysis involving 177 088 patients. British journal of anaesthesia *107*, 659–67.

[Luo et al., 2012] Luo, P., Wang, X. and Tang, X. (2012). Hierarchical face parsing via deep learning. In Computer Vision and Pattern Recognition, IEEE Conference on (CVPR) pp. 2480–2487, Ieee.

[Mallampati et al., 1985] Mallampati, S. R., Gatt, S. P., Gugino, L. D., Desai, S. P., Waraksa, B., Freiberger, D. and Liu, P. L. (1985). A clinical sign to predict difficult tracheal intubation: a prospective study. Canadian Anaesthetists' Society Journal *32*, 429–34.

[Marr and Poggio, 1976] Marr, D. and Poggio, T. (1976). Cooperative Computation of Stereo Disparity. Science *194*, 283–287.

[Marr and Poggio, 1979] Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. Proceedings of the Royal Society London *B 204*, 301–328.

[Martinez et al., 2013] Martinez, B., Valstar, M. F., Binefa, X. and Pantic, M. (2013). Local evidence aggregation for regression-based facial point detection. IEEE Transactions on Pattern Analysis and Machine Intelligence *35*, 1149–1163.

[Martins et al., 2010] Martins, P., Batista, J. and Caseiro, R. (2010). Face Alignment Through 2.5D Active Appearance Models. In Procedings of the British Machine Vision Conference (BMVC) pp. 99.1–99.12,.

[Martins et al., 2013] Martins, P., Caseiro, R. and Batista, J. (2013). Generative face alignment through 2.5D active appearance models. Computer Vision and Image Understanding *117*, 250–268.

[Martins et al., 2012a] Martins, P., Caseiro, R., Henriques, J. and Batista, J. (2012a). Let the Shape Speak-Discriminative Face Alignment using Conjugate Priors. In Proceedings of British Machine Vision Conference (BMVC) pp. 118.1–118.11,.

[Martins et al., 2012b] Martins, P., Caseiro, R., Henriques, J. F. and Batista, J. (2012b). Discriminative Bayesian active shape models. In Proceedings of European Conference on Computer Vision (ECCV) pp. 57–70,.

[Matthews and Baker, 2004] Matthews, I. and Baker, S. (2004). Active Appearance Models Revisited. International Journal of Computer Vision *60*, 135–164.

[Matthews et al., 2007] Matthews, I., Xiao, J. and Baker, S. (2007). 2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting. International Journal of Computer Vision *75*, 93–113.

[Messer et al., 1999] Messer, K., Matas, J., Kittler, J., Luettin, J. and Maitre, G. (1999). XM2VTSDB: The Extended M2VTS Database. In Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99) pp. 1–6,.

[Metzner et al., 2011] Metzner, J., Posner, K. L., Lam, M. S. and Domino, K. B. (2011). Closed claims' analysis. Best practice & research. Clinical anaesthesiology *25*, 263–76.

[Meyer et al., 2002] Meyer, M., Desbrun, M., Schr, P. and Barr, A. H. (2002). Discrete Differential-Geometry Operators for Triangulated 2-Manifolds. Visualization and Mathematics *3*, 52–58.

[Milborrow and Nicolls, 2008] Milborrow, S. and Nicolls, F. (2008). Locating Facial Features with an Extended Active Shape Model. In Proceedings of European Conference on Computer Vision (ECCV) pp. 504–513, Springer Berlin Heidelberg, Berlin, Heidelberg.

[Min et al., 2014] Min, R., Kose, N. and Dugelay, J.-L. (2014). KinectFaceDB: A Kinect Database for Face Recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems *44*, 1534–1548.

[Mitra et al., 2004] Mitra, N. J., Gelfand, N., Pottmann, H. and Guibas, L. (2004). Registration of point cloud data from a geometric optimization perspective. In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing - SGP '04 p. 22, ACM Press, New York, New York, USA.

[Moravec, 1977] Moravec, H. P. (1977). Towards automatic visual obstacle avoidance. In International Conference on Artificial Intelligence (5th: 1977: Massachusetts Institute of Technology).

[Mpiperis et al., 2008] Mpiperis, I., Malassiotis, S. and Strintzis, M. (2008). Bilinear Models for 3-D Face and Facial Expression Recognition. IEEE Transactions on Information Forensics and Security *3*, 498–511.

[Naguib et al., 1999] Naguib, M., Malabarey, T., AlSatli, R. A., Al Damegh, S., Samarkandi, A. H., Damegh, S. A. and Samarkandi, A. H. (1999). Predictive models for difficult laryngoscopy and intubation. A clinical , radiologic and three- dimensional computer imaging study. Canadian Journal of Anesthesiology *46*, 748–759.

[Naguib et al., 2006] Naguib, M., Scamman, F. L., O'Sullivan, C., Aker, J., Ross, A. F., Kosmach, S. and Ensor, J. E. (2006). Predictive performance of three multivariate difficult tracheal intubation models: a double-blind, case-controlled study. Anesthesia and analgesia *102*, 818–24.

[Navarathna et al., 2011] Navarathna, R., Sridharan, S. and Lucey, S. (2011). Fourier Active Appearance Models. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 1919–1926, Ieee.

[Newcombe et al., 2011] Newcombe, R. a., Davison, A. J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D. and Fitzgibbon, A. (2011). KinectFusion: Real-time dense surface mapping and tracking. In 2011 10th IEEE International Symposium on Mixed and Augmented Reality pp. 127–136, IEEE.

[Nguyen and De La Torre, 2008] Nguyen, M. H. and De La Torre, F. (2008). Local minima free parameterized appearance models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Nguyen and De la Torre, 2008] Nguyen, M. H. and De la Torre, F. (2008). Learning image alignment without local minima for face detection and tracking. In Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (FG) pp. 1–7,.

[Nguyen and De La Torre, 2010] Nguyen, M. H. and De La Torre, F. (2010). Metric learning for image alignment. International Journal of Computer Vision *88*, 69–84.

[Nørskov et al., 2015] Nørskov, A. K., Rosenstock, C. V., Wetterslev, J., Astrup, G., Afshari, A. and Lundstrøm, L. H. (2015). Diagnostic accuracy of anaesthesiologists' prediction of difficult airway management in daily clinical practice: a cohort study of 188 064 patients registered in the Danish Anaesthesia Database. Anaesthesia *70*, 272–281.

[Ohtake et al., 2003] Ohtake, Y., Belyaev, A., Alexa, M., Turk, G. and Seidel, H.-P. (2003). Multi-level partition of unity implicits. In ACM SIGGRAPH 2003 Papers on - SIGGRAPH '03 p. 463,.

[Orozco-Díaz et al., 2010] Orozco-Díaz, E., Alvarez-Ríos, J. J., Arceo-Díaz, J. L., Ornelas-Aguirre, J. M., Orozco-Díaz, É., Álvarez-Ríos, J. J., Arceo-Díaz, J. L. and Ornelas-Aguirre, J. M. (2010). Predictive factors of difficult airway with known assessment scales. Cirugia y cirujanos *78*, 393–9.

[Pantic and Rothkrantz, 2000] Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence *22*, 1424–1445.

[Papandreou and Maragos, 2008] Papandreou, G. and Maragos, P. (2008). Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Papert, 1966] Papert, S. (1966). The summer vision project.

[Paquet, 2009] Paquet, U. (2009). Convexity and bayesian constrained local models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) pp. 1193–1199,.

[Parker et al., 1998] Parker, S., Shirley, P., Livnat, Y., Hansen, C. and Sloan, P.-P. (1998). Interactive ray tracing for isosurface rendering. In Proceedings of Visualization pp. 233–238, IEEE.

[Passalis et al., 2005] Passalis, G., Kakadiaris, I., Theoharis, T., Toderici, G. and Murtuza, N. (2005). Evaluation of 3D Face Recognition in the presence of facial expressions: an Annotated Deformable Model approach. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) pp. 171–171,.

[Paysan et al., 2009] Paysan, P., Knothe, R., Amberg, B., Romdhani, S. and Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. In IEEE International Conference on Advanced Video and Signal Based Surveillance pp. 296–301, IEEE.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). Scikit-learn : Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

[Peterson et al., 2005] Peterson, G. N., Domino, K. B., Caplan, R. A., Posner, K. L., Lee, L. A. and Cheney, F. W. (2005). Management of the Difficult Airway: a closed claims analysis. Anesthesiology *103*, 33–39.

[Peyras et al., 2007] Peyras, J., Bartoli, A., Mercier, H. and Dalle, P. (2007). Segmented AAMs Improve Person-Independent Face Fitting. In Proceedings of British Machine Vision Conference (BMVC).

[Phillips et al., 2005] Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J. and Worek, W. (2005). Overview of the Face Recognition Grand Challenge. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–8,.

[Picard, 1995] Picard, R. W. (1995). Affective Computing. Technical Report 321.

[Qu et al., 2015] Qu, C., Gao, H., Monari, E., Beyerer, J. and Thiran, J.-P. (2015). Towards robust cascaded regression for face alignment in the wild. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) pp. 1–9, IEEE.

[Räsänen et al., 2006] Räsänen, J. O., Rosenhouse, G. and Gavriely, N. (2006). Effects of diameter, length, and circuit pressure on sound conductance through endotracheal tubes. IEEE Trans. Biomed. Eng. *53*, 1255–64.

[Ren et al., 2014] Ren, S., Cao, X., Wei, Y. and Sun, J. (2014). Face Alignment at 3000 FPS via Regressing Local Binary Features. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1685–1692, IEEE.

[Ren et al., 2016] Ren, S., Cao, X., Wei, Y. and Sun, J. (2016). Face Alignment via Regressing Local Binary Features. IEEE Transactions on Image Processing *25*, 1233–1245.

[Ringeval et al., 2015] Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D. and Schuller, B. (2015). Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. Pattern Recognition Letters *66*, 22–30.

[Rivera and Martinez, 2012] Rivera, S. and Martinez, A. M. (2012). Learning deformable shape manifolds. Pattern Recognition *45*, 1792–1801.

[Roberts et al., 2007] Roberts, M., Cootes, T. and Adams, J. (2007). Robust Active Appearance Models with Iteratively Rescaled Kernels. In Procedings of the British Machine Vision Conference (BMVC) pp. 17.1–17.10,.

[Roh et al., 2011] Roh, M. C., Oguri, T. and Kanade, T. (2011). Face alignment robust to occlusion. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, (FGW) pp. 239–244,.

[Sagonas et al., 2015] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S. and Pantic, M. (2015). 300 Faces In-The-Wild Challenge: database and results. Image and Vision Computing *47*, 3–18.

[Sagonas et al., 2013a] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S. and Pantic, M. (2013a). A semi-automatic methodology for facial landmark annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) pp. 896–903,.

[Sagonas et al., 2013b] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S. and Pantic, M. (2013b). 300 faces in-the-wild challenge: The first facial landmark Localization Challenge. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 397–403,.

[Sakai et al., 1972] Sakai, T., Nagao, M. and Kanade, T. (1972). Computer analysis and classification of photographs of human faces. Kyoto University.

[Samsoon and Young, 1987] Samsoon, G. L. T. and Young, J. R. B. (1987). Difficult tracheal intubation: a retrospective study. Anaesthesia *42*, 487–490.

[Sánchez-Lozano et al., 2012] Sánchez-Lozano, E., De la Torre, F. and González-Jiménez, D. (2012). Continuous Regression for Non-rigid Image Alignment. In Proceedings of European Conference on Computer Vision (ECCV) pp. 250–263,.

[Sandbach et al., 2012] Sandbach, G., Zafeiriou, S., Pantic, M. and Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. Image and Vision Computing *30*, 683–697.

[Saragih, 2011] Saragih, J. (2011). Principal regression analysis. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2881–2888,.

[Saragih and Göcke, 2009] Saragih, J. and Göcke, R. (2009). Learning AAM fitting through simulation. Pattern Recognition *42*, 2628–2636.

[Saragih and Goecke, 2007] Saragih, J. and Goecke, R. (2007). A nonlinear discriminative approach to AAM fitting. In Proceedings of IEEE International Conference on Computer Vision (ICCV).

[Saragih et al., 2008] Saragih, J. M., Lucey, S. and Cohn, J. F. (2008). Deformable face fitting with soft correspondence constraints. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG).

[Saragih et al., 2009a] Saragih, J. M., Lucey, S. and Cohn, J. F. (2009a). Face Alignment through Subspace Constrained Mean-Shifts. In Proceedings of International Conference on Computer Vision (ICCV).

[Saragih et al., 2009b] Saragih, J. M., Lucey, S. and Cohn, J. F. (2009b). Deformable model fitting with a mixture of local experts. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 2248–2255,.

[Saragih et al., 2009c] Saragih, J. M., Lucey, S. and Cohn, J. F. (2009c). Probabilistic constrained adaptive local displacement experts. In Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW) pp. 288–295,.

[Saragih et al., 2011] Saragih, J. M., Lucey, S. and Cohn, J. F. (2011). Deformable Model Fitting by Regularized Landmark Mean-Shift. International Journal of Computer Vision *91*, 200–215.

[Sauer et al., 2011] Sauer, P., Cootes, T. and Taylor, C. (2011). Accurate regression procedures for active appearance models. In Proceedings of British Machine Vision Conference (BMVC) pp. 1–11,.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklioglu, H., Çeliktutan, O. and Gökberk, B. (2008). Bosphorus Database for 3D Face Analysis. In Workshop on Biometrics and Identity Management (BIOID) pp. 47–56,.

[Schendel and Hatcher, 2010] Schendel, S. A. and Hatcher, D. (2010). Automated 3-dimensional airway analysis from cone-beam computed tomography data. Journal of oral and maxillofacial surgery : official journal of the American Association of Oral and Maxillofacial Surgeons *68*, 696–701.

[Schoettker et al., 2014] Schoettker, P., Cuendet, G., Perruchoud, C., Sorci, M. and Thiran, J.-P. (2014). Difficult intubation or ventilation or extubation prediction system. U.S. Patent Application No 15/027,899.

[Serocki et al., 2010] Serocki, G., Bein, B., Scholz, J. and Dörges, V. (2010). Management of the predicted difficult airway: A comparison of conventional blade laryngoscopy with video-assisted blade laryngoscopy and the glidescope. European Journal of Anaesthesiology *27*, 24–30.

[Shen et al., 2013] Shen, X., Lin, Z., Brandt, J. and Wu, Y. (2013). Detecting and aligning faces by image retrieval. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3460–3467,.

[Shiga et al., 2005] Shiga, T., Wajima, Z., Inoue, T. and Sakamoto, A. (2005). Predicting Difficult Intubation in Apparently Normal Patients: A Meta-analysis of Bedside Screening Test Performance. Anesthesiology *103*, 429–437.

[Shiyang Cheng et al., 2015] Shiyang Cheng, Marras, I., Zafeiriou, S. and Pantic, M. (2015). Active nonrigid ICP algorithm. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) pp. 1–8, IEEE.

[Smith et al., 2013] Smith, B. M., Zhang, L., Brandt, J., Lin, Z. and Yang, J. (2013). Exemplar-based face parsing. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3484–3491,.

[Sorkine, 2009] Sorkine, O. (2009). Least-squares rigid motion using svd. Technical Report February.

[Sozou et al., 1995] Sozou, P., Cootes, T., Taylor, C. and Di Mauro, E. (1995). Non-linear generalization of point distribution models using polynomial regression. Image and Vision Computing *13*, 451–457.

[Sozou et al., 1997] Sozou, P. D., Cootes, T., Taylor, C., Di Mauro, E. and Lanitis, A. (1997). Non-linear point distribution modelling using a multi-layer perceptron. Image and Vision Computing *15*, 457–463.

[Sukno et al., 2007] Sukno, F. M., Ordás, S., Butakoff, C., Cruz, S. and Frangi, A. F. (2007). Active shape models with invariant optimal features: Application to facial analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence *29*, 1105–1117.

[Sumner and Popović, 2004] Sumner, R. W. and Popović, J. (2004). Deformation transfer for triangle meshes. ACM Transactions on Graphics *23*, 399.

[Sumner et al., 2007] Sumner, R. W., Schmid, J. and Pauly, M. (2007). Embedded deformation for shape manipulation. ACM Transactions on Graphics *26*, 80.

[Sun et al., 2013] Sun, Y., Wang, X. and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3476–3483,.

[Sung et al., 2007] Sung, J., Kanade, T. and Kim, D. (2007). A unified gradient-based approach for combining ASM into AAM. International Journal of Computer Vision *75*, 297–309.

[Sung et al., 2008] Sung, J., Kanade, T. and Kim, D. (2008). Pose robust face tracking by combining active appearance models and cylinder head models. International Journal of Computer Vision *80*, 260–274.

[Suzuki et al., 2007] Suzuki, N., Isono, S., Ishikawa, T., Kitamura, Y., Takai, Y. and Nishino, T. (2007). Submandible Angle in Nonobese Patients with Difficult Tracheal Intubation. Anesthesiology *106*, 916–923.

[Szeliski, 2011] Szeliski, R. (2011). Computer Vision: Algorithms and Applications. Texts in Computer Science, Springer, London.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 1701–1708, IEEE.

[Tam et al., 2013] Tam, G. K. L., Zhi-Quan Cheng, Yu-Kun Lai, Langbein, F. C., Yonghuai Liu, Marshall, D., Martin, R. R., Xian-Fang Sun and Rosin, P. L. (2013). Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid. IEEE Transactions on Visualization and Computer Graphics *19*, 1199–1217.

[Taubin, 1995] Taubin, G. (1995). A signal processing approach to fair surface design. In SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques pp. 351–358, ACM Press, New York, New York, USA.

[Tena et al., 2006] Tena, J., Hamouz, M., Hilton, A. and Illingworth, J. (2006). A Validated Method for Dense Non-rigid 3D Face Registration. In 2006 IEEE International Conference on Video and Signal Based Surveillance pp. 81–81, IEEE.

[Teoh et al., 2010] Teoh, W. H. L., Saxena, S., Shah, M. K. and Sia, a. T. H. (2010). Comparison of three videolaryngoscopes: Pentax Airway Scope, C-MAC, Glidescope vs the Macintosh laryngoscope for tracheal intubation. Anaesthesia *65*, 1126–32.

[Thies et al., 2016] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and NieBner, M. (2016). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2387–2395, IEEE.

[Toderici et al., 2014] Toderici, G., Evangelopoulos, G., Fang, T., Theoharis, T. and Kakadiaris, I. A. (2014). UHDB11 database for 3D-2D face recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) *8333 LNCS*, 73–86.

[Tomasi and Manduchi, 1998] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In Proceedings of the International Conference on Computer Vision (ICCV) pp. 839–846, Narosa Publishing House.

[Tresadern et al., 2010] Tresadern, P., Sauer, P. and Cootes, T. F. (2010). Additive Update Predictors in Active Appearance Models. In Procedings of the British Machine Vision Conference (BMVC) pp. 91.1–91.12, British Machine Vision Association.

[Tresadern et al., 2009] Tresadern, P. A., Bhaskar, H., Adeshina, S. A., Taylor, C. J. and Cootes, T. F. (2009). Combining Local and Global Shape Models for Deformable Object Matching. In Procedings of the British Machine Vision Conference (BMVC) pp. 95.1–95.12, British Machine Vision Association.

[Tresadern et al., 2012] Tresadern, P. a., Ionita, M. C. and Cootes, T. F. (2012). Real-Time Facial Feature Tracking on a Mobile Device. International Journal of Computer Vision *96*, 280–289.

[Tzimiropoulos et al., 2012] Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S. and Pantic, M. (2012). Generic Active Appearance Models Revisited. In Proceedings of Asian Conference on Computer Vision (ACCV) pp. 650–663,.

[Tzimiropoulos and Pantic, 2013] Tzimiropoulos, G. and Pantic, M. (2013). Optimization problems for fast AAM fitting in-the-wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 593–600,.

[Uricar et al., 2012] Uricar, M., Franc, V., Hlavac, V. and Hlavác, V. (2012). Detector of facial landmarks learned by the structured output SVM. In Proceedings of the 7th International Conference on Computer Vision Theory and Applications pp. 547–556,.

[Vallet and Lévy, 2008] Vallet, B. and Lévy, B. (2008). Spectral Geometry Processing with Manifold Harmonics. Computer Graphics Forum *27*, 251–260.

[Valstar et al., 2010] Valstar, M., Martinez, B., Binefa, X. and Pantic, M. (2010). Facial point detection using boosted regression and graph models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2729–2736,.

[Valstar et al., 2015] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M. and Cohn, J. F. (2015). FERA 2015 - second Facial Expression Recognition and Analysis challenge. In Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) vol. 06, pp. 1–8,.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579–2605.

[van Kaick et al., 2011] van Kaick, O., Zhang, H., Hamarneh, G. and Cohen-Or, D. (2011). A Survey on Shape Correspondence. Computer Graphics Forum *30*, 1681–1707.

[Vetter and Blanz, 1998] Vetter, T. and Blanz, V. (1998). Estimating coloured 3d face models from single images: An example based approach. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) *1407*, 499–513.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 511–518, IEEE.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. International Journal of Computer Vision *57*, 137–154.

[Vlasic et al., 2005] Vlasic, D., Brand, M., Pfister, H. and Popović, J. (2005). Face transfer with multilinear models. ACM Transactions on Graphics *24*, 426.

[Vogler et al., 2007] Vogler, C., Li, Z., Kanaujia, A., Goldenstein, S. and Metaxas, D. (2007). The Best of Both Worlds: Combining 3D Deformable Models with Active Shape Models. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 1–7,.

[Vukadinovic and Pantic, 2005] Vukadinovic, D. and Pantic, M. (2005). Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. In Proceedings of IEEE International Conference on Systems, Man and Cybernetics pp. 1692–1698,.

[Wang et al., 2014] Wang, N., Gao, X., Tao, D. and Li, X. (2014). Facial Feature Point Detection: A Comprehensive Survey.

[Wang et al., 2008a] Wang, Y., Lucey, S. and Cohn, J. F. (2008a). Enforcing convexity for improved alignment with constrained local models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Wang et al., 2008b] Wang, Y., Lucey, S., Cohn, J. F. and Saragih, J. (2008b). Non-rigid face tracking with local appearance consistency constraint. In Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (FG) pp. 1–8, Ieee.

[Weise et al., 2011] Weise, T., Bouaziz, S., Li, H. and Pauly, M. (2011). Realtime performance-based facial animation. ACM Transactions on Graphics, SIGGRAPH 2011 *30*, 1–9.

[Weise et al., 2007] Weise, T., Leibe, B. and Van Gool, L. (2007). Fast 3D Scanning with Automatic Motion Compensation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–8, IEEE.

[Weise et al., 2009] Weise, T., Li, H., Van Gool, L. and Pauly, M. (2009). Face / Off : Live Facial Puppetry. In Eurographics/ACM SIGGRAPH Symposium on Computer Animation pp. 7–16,.

[Wilson et al., 1988] Wilson, M. E., Spiegelhalter, D., Robertson, J. A. and Lesser, P. (1988). Predicting Difficult Intubation. British journal of anaesthesia *61*, 211–216.

[Wimmer et al., 2008] Wimmer, M., Stulp, F., Pietzsch, S. and Radig, B. (2008). Learning local objective functions for robust face model fitting. IEEE Transactions on Pattern Analysis and Machine Intelligence *30*, 1357–1370.

[Wu et al., 2013] Wu, Y., Wang, Z. and Ji, Q. (2013). Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3452–3459,.

[Xiong and De la Torre, 2013] Xiong, X. and De la Torre, F. (2013). Supervised Descent Method and Its Applications to Face Alignment. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 532–539, IEEE.

[Xiong and De la Torre, 2015] Xiong, X. and De la Torre, F. (2015). Global supervised descent method. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2664–2673, IEEE.

[Xiong and Torre, 2014] Xiong, X. and Torre, F. D. (2014). Supervised Descent Method for Solving Nonlinear Least Squares Problems in Computer Vision. In arXiv preprint arXiv:1405.0601 pp. 1–15,.

[Yang and Patras, 2013a] Yang, H. and Patras, I. (2013a). Face parts localization using structured-output regression forests. In Proceedings of Asian Conference on Computer Vision (ACCV) pp. 667–679,.

[Yang and Patras, 2013b] Yang, H. and Patras, I. (2013b). Sieving regression forest votes for facial feature detection in the wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 1936–1943,.

[Yang et al., 2002] Yang, M.-h., Kriegman, D. J. and Ahuja, N. (2002). Detecting Faces in Images : A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence *24*, 34–58.

[Yang and Ramanan, 2011] Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In Computer Vision and Pattern Recognition, IEEE Conference on (CVPR) pp. 1385–1392,.

[Yang and Ramanan, 2013] Yang, Y. and Ramanan, D. (2013). Articulated Human Detection with Flexible Mixtures-of-Parts. IEEE Transactions on Pattern Analysis and Machine Intelligence *35*, 1–15.

[Yang Wang et al., 2009] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang and Samaras, D. (2009). Face Relighting from a Single Image under Arbitrary Unknown Lighting Conditions. IEEE Transactions on Pattern Analysis and Machine Intelligence *31*, 1968–1984.

[Yentis, 2002] Yentis, S. M. (2002). Predicting difficult intubation - worthwhile exercise or pointless ritual? Anesthesia *57*, 105–109.

[Yentis and Lee, 1998] Yentis, S. M. and Lee, D. J. H. (1998). Evaluation of an improved scoring system for the grading of direct laryngoscopy. Anaesthesia *53*, 1041–4.

[Yin et al., 2008] Yin, L., Chen, X., Sun, Y., Worm, T. and Reale, M. (2008). A high-resolution 3D dynamic facial expression database. In 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition number 1 pp. 1–6,.

[Yin et al., 2006] Yin, L., Wei, X., Sun, Y., Wang, J. and Rosato, M. J. (2006). A 3D Facial Expression Database For Facial Behavior Research. In 7th International Conference on Automatic Face and Gesture Recognition (FGR06) pp. 211–216, Ieee.

[Yu et al., 2013] Yu, X., Huang, J., Zhang, S., Yan, W. and Metaxas, D. N. (2013). Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model. In Proceedings of IEEE International Conference on Computer Vision (ICCV) pp. 1944–1951,.

[Yüce et al., 2013] Yüce, A., Arar, N. M. and Thiran, J. P. (2013). Multiple local curvature gabor binary patterns for facial action recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) *8212 LNCS*, 136–147.

[Zafeiriou et al., 2011] Zafeiriou, S., Hansen, M., Atkinson, G., Argyriou, V., Petrou, M., Smith, M. and Smith, L. (2011). The Photoface database. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) pp. 132–139, IEEE.

[Zafeiriou et al., 2015] Zafeiriou, S., Zhang, C. and Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding *138*, 1–24.

[Zell and Botsch, 2013] Zell, E. and Botsch, M. (2013). ElastiFace. In Proceedings of the Symposium on Non-Photorealistic Animation and Rendering - NPAR '13 p. 15, ACM Press, New York, New York, USA.

[Zhang and Zhang, 2010] Zhang, C. and Zhang, Z. (2010). A Survey of Recent Advances in Face Detection. Technical Report June.

[Zhang et al., 2010] Zhang, H., Van Kaick, O. and Dyer, R. (2010). Spectral Mesh Processing. Computer Graphics Forum *29*, 1865–1894.

[Zhang et al., 2004] Zhang, L., Snavely, N., Curless, B. and Seitz, S. M. (2004). Spacetime Faces: High Resolution Capture for Modeling and Animation. In Proceedings SIGGRAPH pp. 548–558,.

[Zhang et al., 2017] Zhang, L., Zhang, D., Sun, M.-m. and Chen, F.-m. (2017). Facial beauty analysis based on geometric feature: Toward attractiveness assessment application. Expert Systems with Applications *82*, 252–265.

[Zhang et al., 2013] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A. and Liu, P. (2013). A High-Resolution Spontaneous 3D Dynamic Facial Expression Database. In 2013 IEEE International Conference on Automatic Face & Gesture Recognition number 1 pp. 1–6,.

[Zhang et al., 2014] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P. and Girard, J. M. (2014). BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. Image and Vision Computing *32*, 692–706.

[Zhao et al., 2013] Zhao, Q., Rosenbaum, K., Okada, K., Zand, D. J., Sze, R., Summar, M. and Linguraru, M. G. (2013). Automated down syndrome detection using facial photographs. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS vol. 2013, pp. 3670–3,.

[Zhao et al., 2003] Zhao, W., Chellappa, R., Phillips, P. and Rosenfeld, a. (2003). Face recognition: A literature survey. Acm Computing Surveys *35*, 399–458.

[Zhao et al., 2013] Zhao, X., Shan, S., Chai, X. and Chen, X. (2013). Cascaded shape space pruning for robust facial landmark detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 1033–1040,.

[Zhili Mao et al., 2004] Zhili Mao, Siebert, J., Cockshott, W. and Ayoub, A. (2004). Constructing dense correspondences to analyze 3D facial change. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. pp. 144–148 Vol.3, IEEE.

[Zhong et al., 2007] Zhong, C., Sun, Z. and Tan, T. (2007). Robust 3D Face Recognition Using Learned Visual Codebook. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–6, IEEE.

[Zhou and Comaniciu, 2007] Zhou, S. K. and Comaniciu, D. (2007). Shape Regression Machine. In Information Processing in Medical Imaging vol. 20, pp. 13–25. Springer Berlin Heidelberg Berlin, Heidelberg.

[Zhou et al., 2003] Zhou, Y., Gu, L. and Zhang, H.-J. (2003). Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 109–116,.

[Zhu and Ramanan, 2012] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2879–2886, Ieee.

[Zimmermann et al., 2016] Zimmermann, M., Mehdipour Ghazi, M., Ekenel, H. K. and Thiran, J.-P. (2016). Visual Speech Recognition Using PCA Networks and LSTMs in a Tandem GMM-HMM System. In Proceedings of Asian Conference on Computer Vision- Workshop Multi-view Lip-reading Challenge (ACCVW).

Gabriel **Cuendet**

Post-doctoral researcher · Computer Vision & Machine Learning

Place du Petit-St-Jean 13, 1700 Fribourg, Switzerland

gabriel.cuendet@alumni.epfl.ch | gcuendet.github.io | gcuendet | gcuendet | gcuendet

Education

Ph.D., Electrical Engineering — *Lausanne, Switzerland*

Ecole Polytechnique Fédérale de Lausanne (EPFL) — *August 2017*

- Thesis Topic: *Towards 3D facial morphometry: facial image analysis applications in anesthesiology and 3D spectral nonrigid registration*
- Adviser: Prof. Jean-Philippe Thiran

M.S., Electrical Engineering **GPA: 5.56 (6.0 scale)** — *Lausanne, Switzerland*

Ecole Polytechnique Fédérale de Lausanne (EPFL) — *July 2012*

- Thesis Topic: Thesis Topic: *Difficult Intubation Assessment from Video*
- Area of Study: Major in **information technologies** and minor in **biomedical technologies**

Professional Experience

Ecole Polytechnique Fédérale de Lausanne (EPFL) — *Lausanne, Switzerland*

Research assistant — *September 2012 to present*

Objective: Automatically predict difficulty of intubation and develop a new 3D face model

Mission: Conduct research in collaboration with CHUV and nViso, collect data in hospitals, develop a C++ library for facial images analysis, record and align a 3D database of faces, supervise students in projects related to facial images analysis

Technologies: C++, Python, Face Alignment (AAM, CLM, SDM, LBF), Machine Learning, 3D Geometry, Spectral Mesh Processing, 3D Face Models

Results: EU Patent application, scientific publications

Teaching assistant — *September 2008 to June 2011*

Teaching Assistant for the courses and labs: Introduction to electrical engineering, Measurement Systems, Programming (C++)

IBM Research — *Zürich, Switzerland*

Research intern — *September 2015 to February 2016*

Objective: Automatically extract numerical data from scientific charts images

Mission: Conduct research, collect and organize data, develop and test code, write a scientific article and a patent application

Technologies: C++, Python, Image Processing, Machine Learning, Markov Logic Network

Results: US Patent application, conference article submission, post-doc position opening to continue the project

ABB, Corporate Research Center — *Bangalore, India*

Intern — *July 2010 to September 2010*

Objective: Reduce the use of big temporary objects at execution time in order to achieve real-time simulation of electrical systems

Mission: Performed simulations and explored advanced concepts of C++

Technologies: C++, expression templates, template meta-programming

Results: Internship report containing preliminary results

Skills

Programming Languages

C++, Python, OpenCV library, CMake, Scikit-learn and NumPy libraries, Matlab, Bash, TeX (LaTeX, BibTeX)

French: mother tongue

English: Excellent knowledge (professional language since 2010)

Swedish: Good knowledge (exchange year in Sweden, 2002-2003)

German: School knowledge (9 years courses)

Awards

Institute for Pure & Applied Mathematics (IPAM), UCLA *Los Angeles, USA*

Full grant for attending the Graduate Summer School: Computer Vision *Summer 2013*

Peer-reviewed Publications

Refereed Journal Publications

[1] **G. L. Cuendet**, C. Ecabert, M. Zimmermann, H. K. Ekenel, J.-P. Thiran. 3D Spectral Nonrigid Registration of Facial Expression Scans. *submitted to IEEE Transactions on Visualization and Computer Graphics*, April 2017

[2] A. Yüce, H. Gao, **G. L. Cuendet**, J.-P. Thiran. Action Units and Their Cross-Correlations for Prediction of Cognitive Load during Driving. *IEEE Transactions on Affective Computing*, Jun. 2016
doi:10.1109/TAFFC.2016.2584042

[3] **G. L. Cuendet**, P. Schoettker, A. Yüce, M. Sorci, H. Gao, C. Perruchoud, and J.-P. Thiran. Facial image analysis for fully automatic prediction of difficult endotracheal intubation. *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 328-339, Feb. 2016.
doi:10.1109/TBME.2015.2457032

Conference Publications

[1] A. Yüce, J.-P. Thiran, M. Sorci, P. Schoettker and C. Perruchoud. Automatic Mallampati Classification Using Active Appearance Models. **G. L. Cuendet**, A. Yüce, J.-P. Thiran, M. Sorci, P. Schoettker, C. Perruchoud. Automatic Mallampati Classification Using Active Appearance Models. *ICPR International Workshop on Pattern Recognition for Healthcare Analytics*, 2012.

Patents

[1] **G. L. Cuendet**, P. Staar, M. Gabrani and K. Bekas. A method and a system to fully-automatically and quantitatively analyze technical diagrams. Patent to be filed at the US Patent Office.

[2] P. Schoettker, **G. L. Cuendet**, C. Perruchoud, M. Sorci and J.-P. Thiran. Difficult intubation or ventilation prediction system. Patent pending at the European Patent Office, October 2013.

Conference Abstracts

[1] P. Schoettker, **G. L. Cuendet**, J.-P. Thiran, M. Sorci, C. Perruchoud. Automated assessment of difficult ventilation with facial recognition techniques. *Swiss Medical Weekly, 143(201), p. 4*, 2013.

[2] P. Schoettker, **G. L. Cuendet**, J.-P. Thiran, M. Sorci, A. Yüce, C. Perruchoud. Automatic Prediction of Difficult Intubation from Video. *European Journal of Anaesthesiology (EJA), p. 269*, 2013

[3] P. Schoettker, **G. L. Cuendet**, J.-P. Thiran, M. Sorci, A. Yüce and C. Perruchoud. Automatic Assessment of Difficult Intubation from Video. *Swiss Medical Forum*, vol. 42(35), pp. 4-5, 2012.

[4] F. I. Karahanoglu, **G. L. Cuendet**, J. Britz, D. V. D. Ville and C. Michel. Multidimensional Random Walk Embedding to Reveal the EEG Microstates Dynamics. *Proceedings of the Seventh Alpine Brain Imaging Meeting*, p.65, 2012.

Extra-curricular

Certificat amateur de violon (certificate of violin amateur studies) *Fribourg, Switzerland*

Conservatoire de Fribourg *June 2009*

Chamber music **2009 to present**

- Violinist of the ”Chromatique” piano trio. We perform public concerts in the french speaking part of Switzerland, playing the classical and romantic repertoire.
- Chamber music master classes in Switzerland and Germany with amongst others: the Mandelring quartet, Paul Cocker, Joel Marosi or the Trio Lenitas.

Orchestra musician (OSUL) **2012 to present**

- Violinist in the Lausanne symphonic university orchestra. The orchestra gives 3 concerts per year and plays the romantic and modern repertoire for large symphonic orchestra.

References

Available upon request